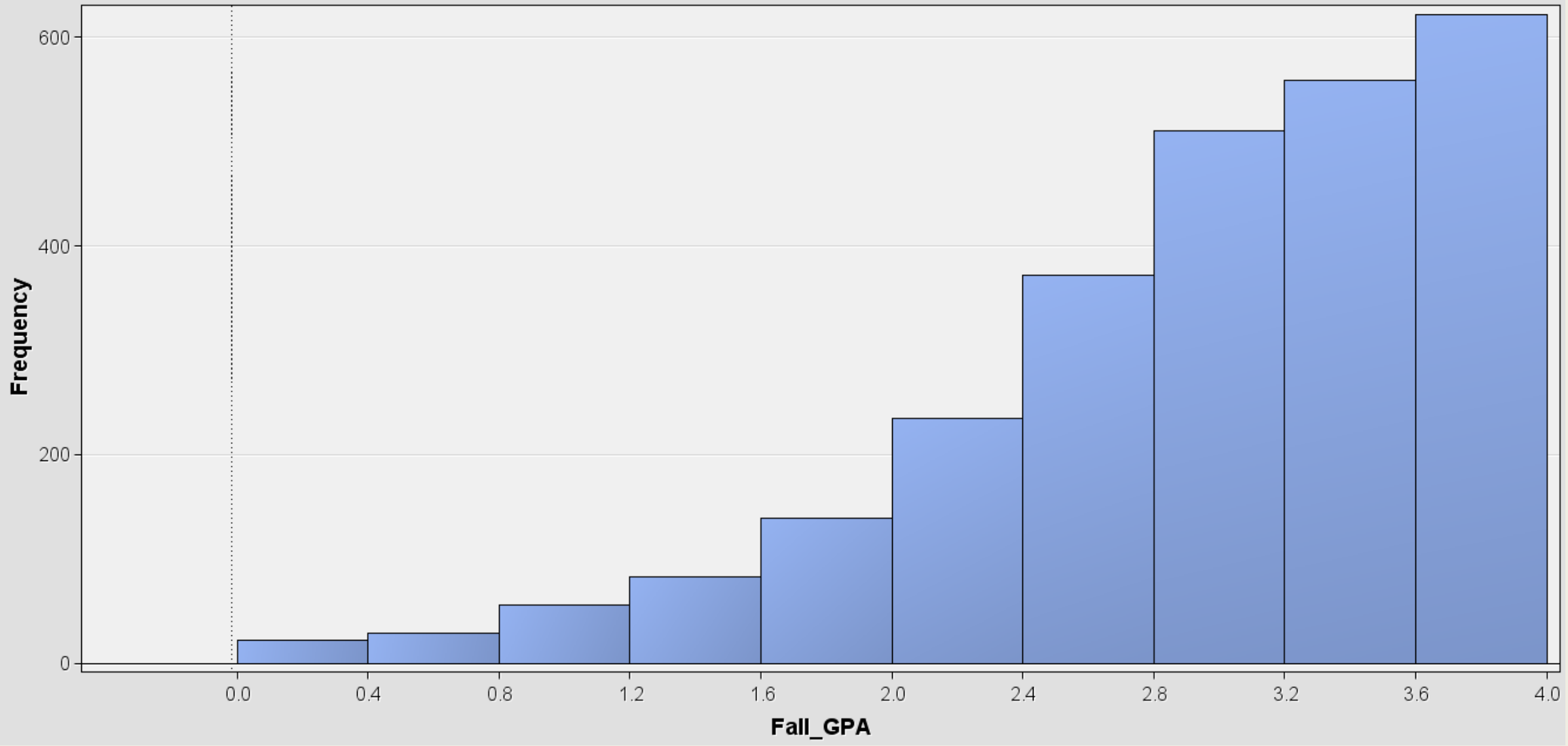
# WORKSHOP 5

By Group 5-
Anand Mohan Thakur
Josh Shaji
Poonam Bhaliyan
Poornima Singh
Prateek Ramjanam Singh

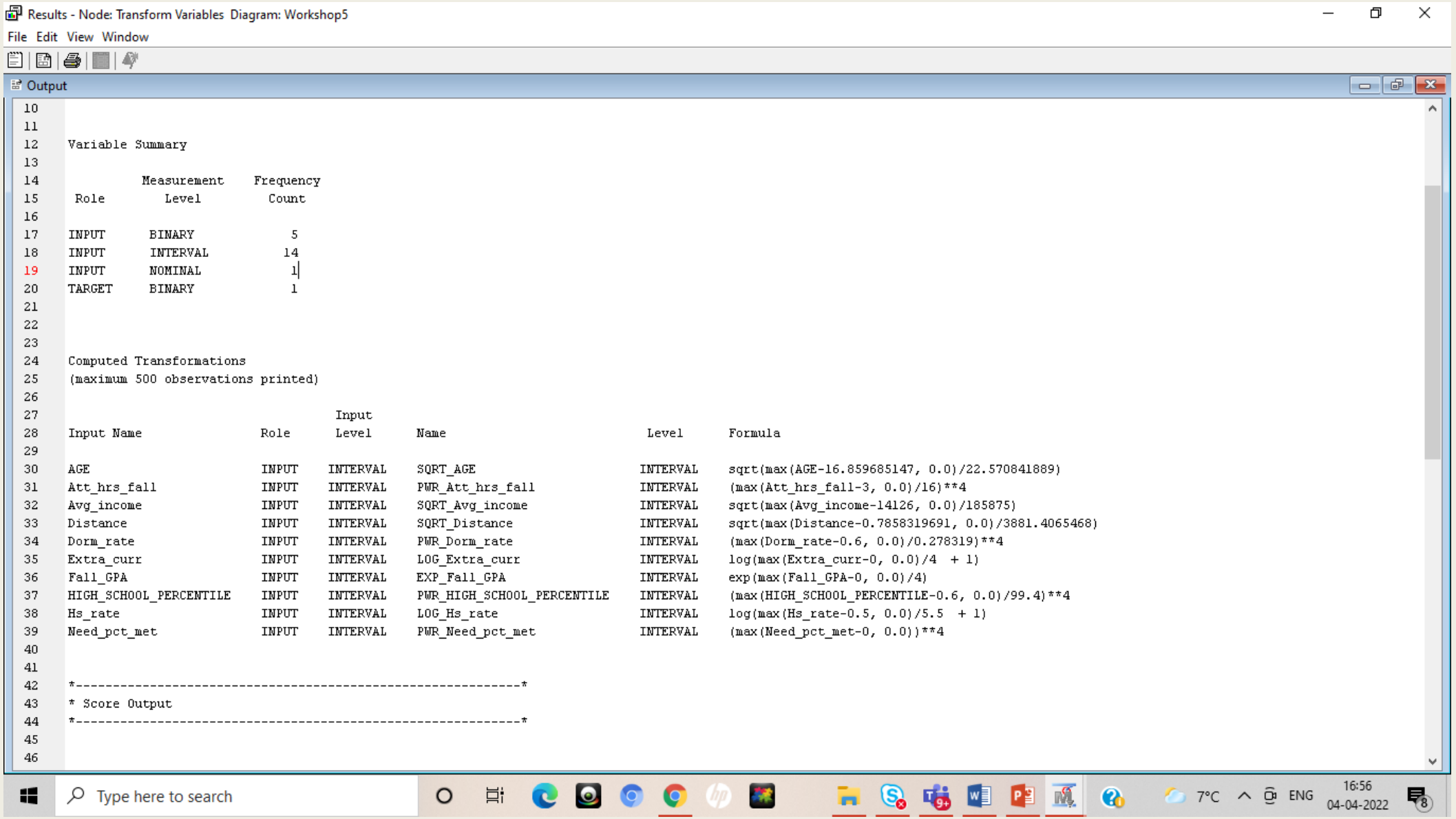
# PART II: LOGISTIC REGRESSION

# Answer 3: Step 2 - The Fall_GPA Plot



# Answer 3: Step 5 – Transformation of Fall_GPA

The formula used for transformation of Fall_GPA is: **exp(max(Fall_GPA-0, 0.0)/4)**
The skewness for Fall_GPA is -0.99401 whereas for the transformed Fall_GPA is -0.49838




Variable Summary

| Role | Measurement Level | Frequency Count |
|---|---|---|
| INPUT | BINARY | 5 |
| INPUT | INTERVAL | 14 |
| INPUT | NOMINAL | 1 |
| TARGET | BINARY | 1 |

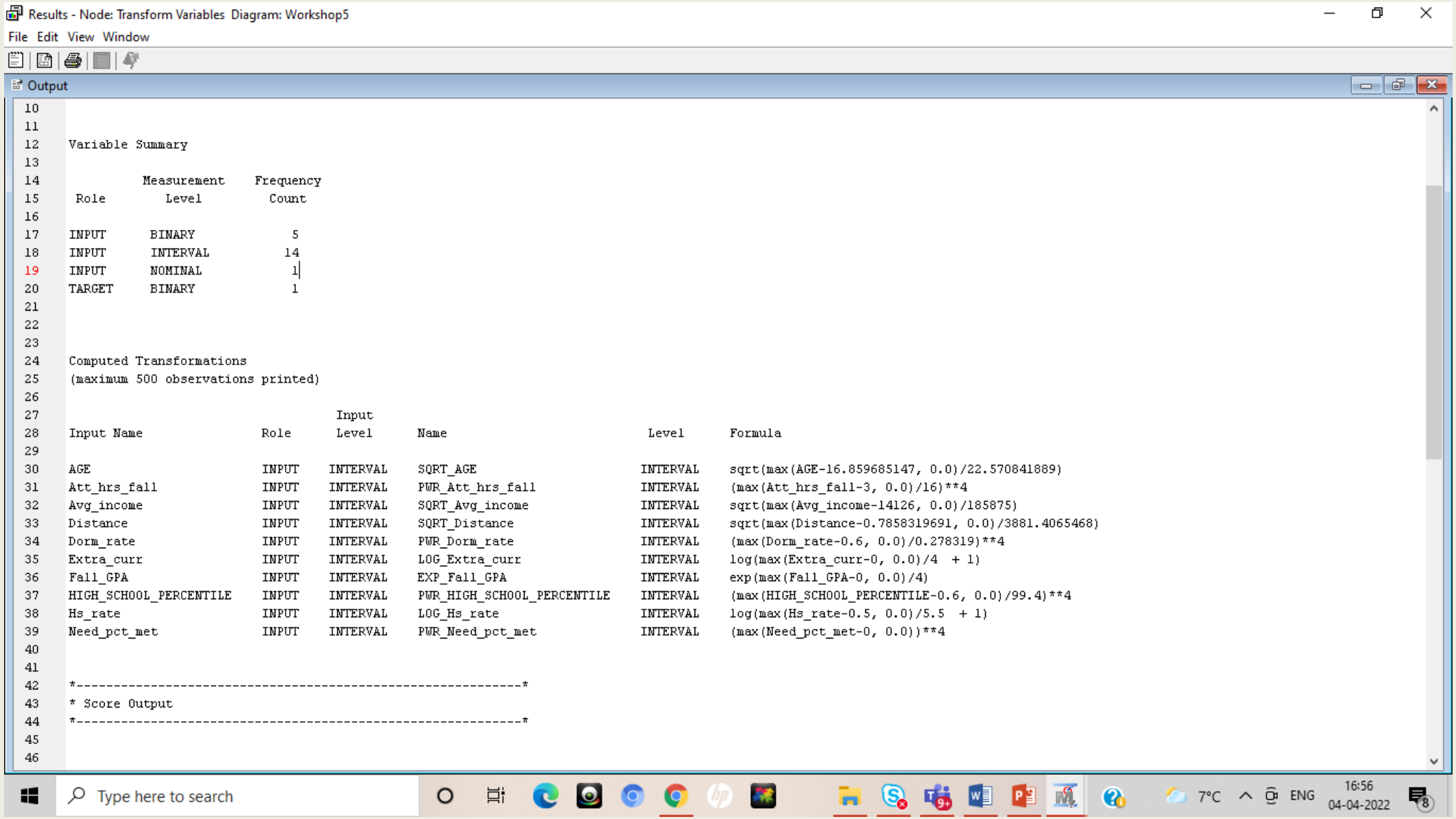
Computed Transformations
(maximum 500 observations printed)

| Input Name | Role | Input Level | Name | Level | Formula |
|---|---|---|---|---|---|
| AGE | INPUT | INTERVAL | SQRT_AGE | INTERVAL | sqrt(max(AGE-16.859685147, 0.0)/22.570841889) |
| Att_hrs_fall | INPUT | INTERVAL | PWR_Att_hrs_fall | INTERVAL | (max(Att_hrs_fall-3, 0.0)/16)**4 |
| Avg_income | INPUT | INTERVAL | SQRT_Avg_income | INTERVAL | sqrt(max(Avg_income-14126, 0.0)/185875) |
| Distance | INPUT | INTERVAL | SQRT_Distance | INTERVAL | sqrt(max(Distance-0.7858319691, 0.0)/3881.4065468) |
| Dorm_rate | INPUT | INTERVAL | PWR_Dorm_rate | INTERVAL | (max(Dorm_rate-0.6, 0.0)/0.278319)**4 |
| Extra_curr | INPUT | INTERVAL | LOG_Extra_curr | INTERVAL | log(max(Extra_curr-0, 0.0)/4 + 1) |
| Fall_GPA | INPUT | INTERVAL | EXP_Fall_GPA | INTERVAL | exp(max(Fall_GPA-0, 0.0)/4) |
| HIGH_SCHOOL_PERCENTILE | INPUT | INTERVAL | PWR_HIGH_SCHOOL_PERCENTILE | INTERVAL | (max(HIGH_SCHOOL_PERCENTILE-0.6, 0.0)/99.4)**4 |
| Hs_rate | INPUT | INTERVAL | LOG_Hs_rate | INTERVAL | log(max(Hs_rate-0.5, 0.0)/5.5 + 1) |
| Need_pct_met | INPUT | INTERVAL | PWR_Need_pct_met | INTERVAL | (max(Need_pct_met-0, 0.0))**4 |

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

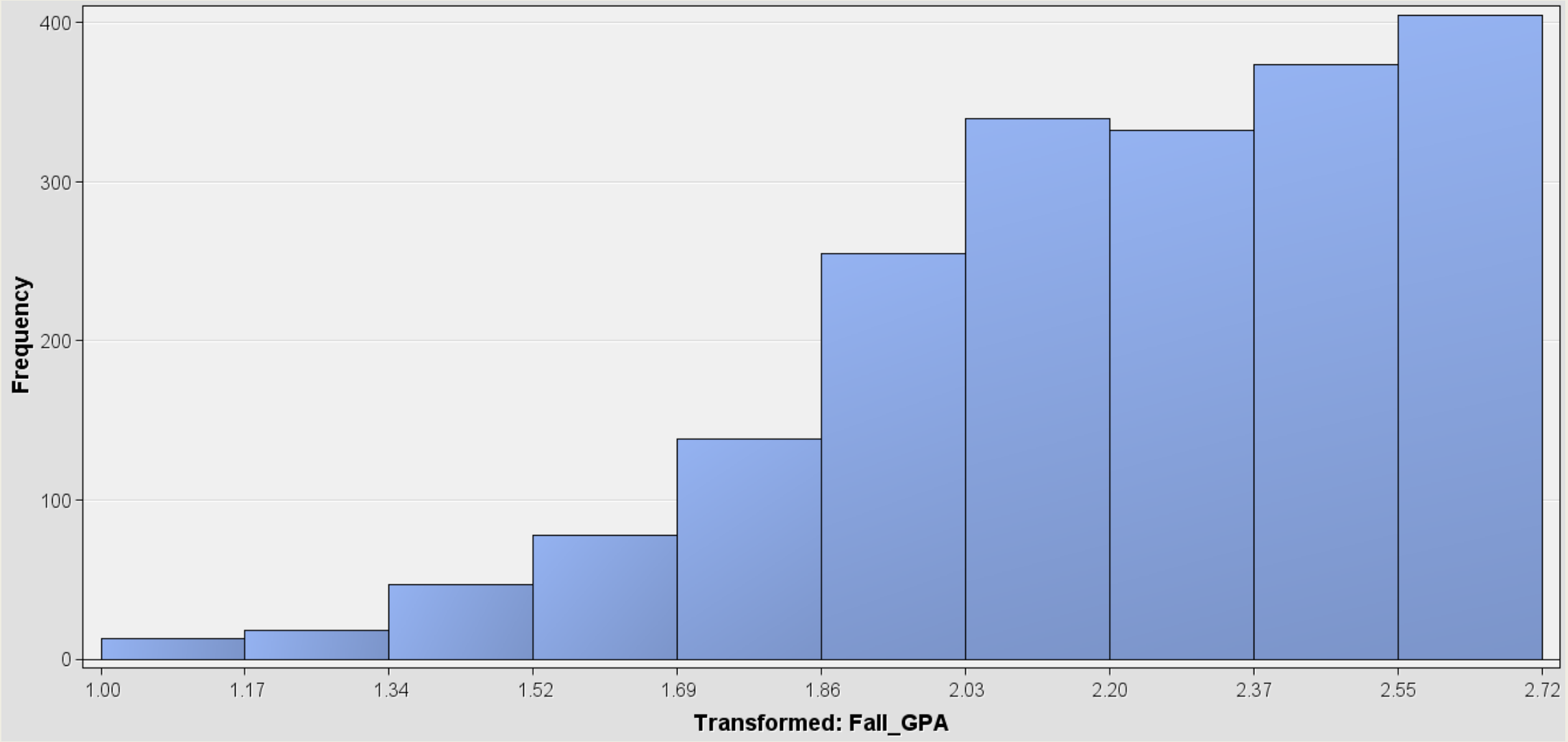
# Answer 3: Step 5 – Transformation of Distance

The formula used for transformation of Distance is: **sqrt(max(Distance-0.7858319691, 0.0)/3881.4065468)**
The skewness for Distance is 3.748368 whereas for the transformed Distance is 1.755574. The skewness has reduced



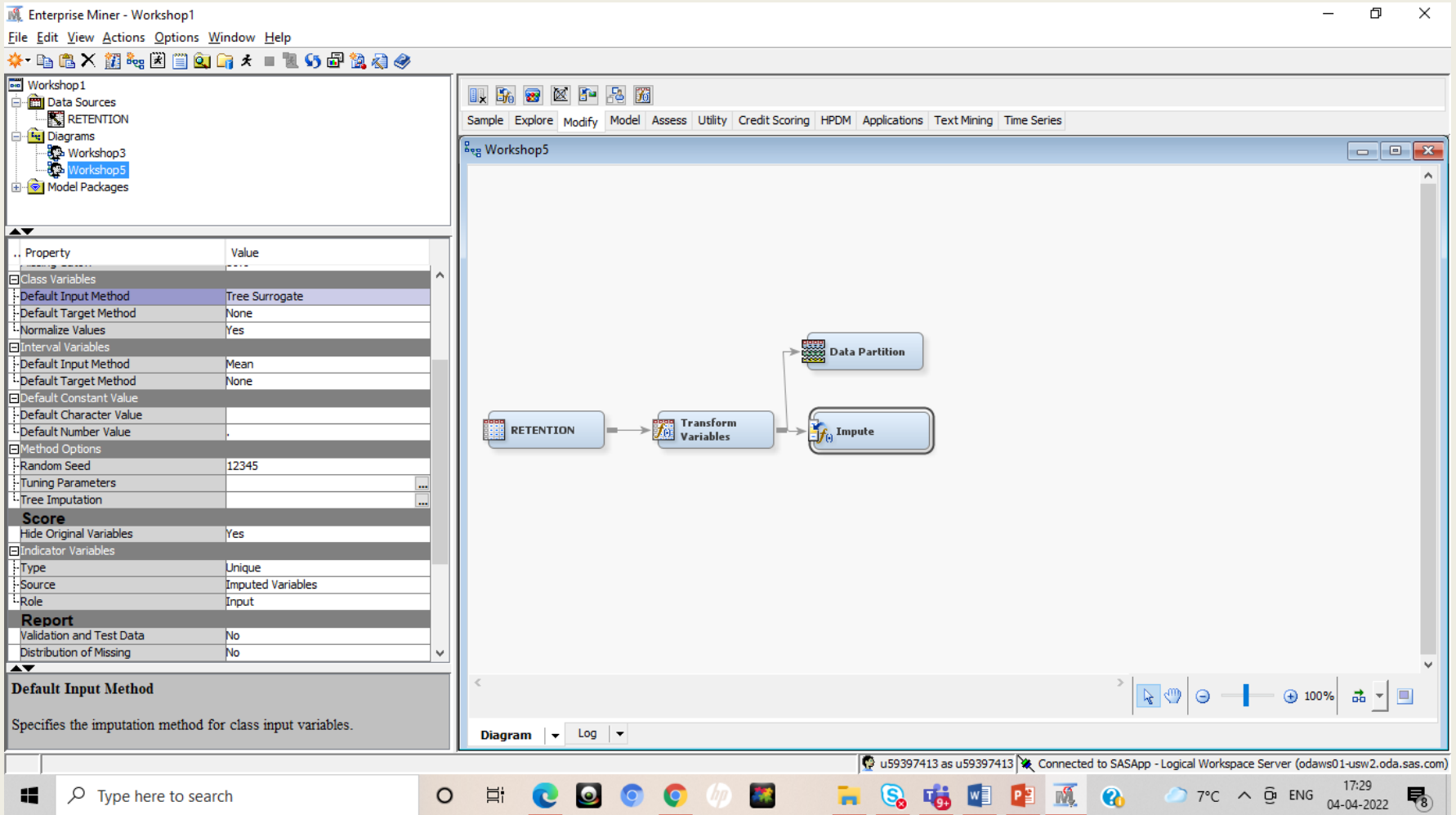
# Answer 3: Step 7 – Histogram of EXP Fall_GPA

In the previous graph of Fall_GPA the values range from 0 to 4, whereas in the below graph with exponential values of Fall_GPA, the values range from 1 to 2.72, making the tail less flatter



# Answer 3: Step 9 – Tree Surrogate

**Tree Surrogate** — The Tree Surrogate setting is used to replace missing interval variable values by using the same algorithm as Tree Imputation (in Tree Imputation Use the Tree setting to replace missing interval variable values with replacement values that are estimated by analysing each input as a target. The remaining input and rejected variables are used as predictors), except with the addition of surrogate splitting rules. A surrogate rule is a backup to the main splitting rule. When the main splitting rule relies on an input whose value is missing, the next surrogate is invoked. If missing values prevent the main rule and all the surrogates from applying to an observation, the main rule assigns the observation to the branch that is assigned to receive missing values.

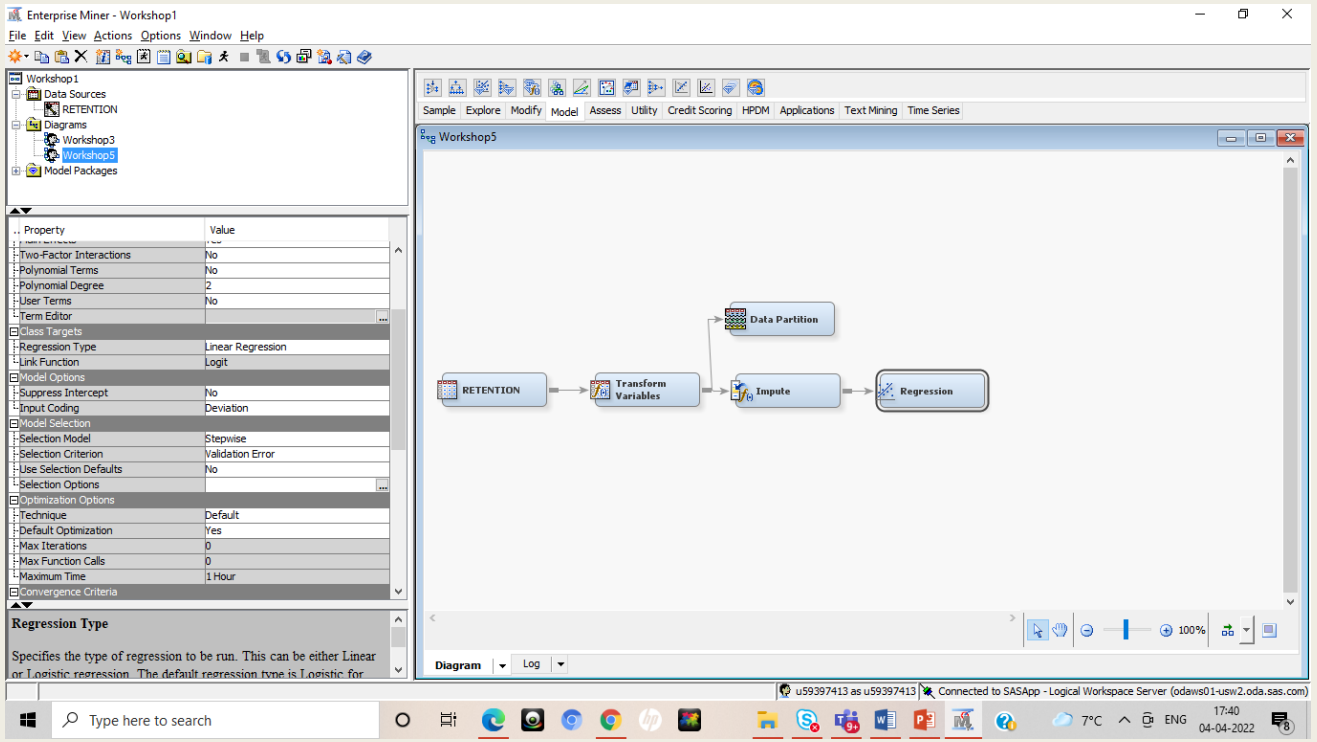


# Answer 3: Step 11 – Stepwise Regression

**Stepwise Regression is a variable selection method.**

**Stepwise**: In Stepwise Regression selection begins, by default, with no candidate effects in the model and then systematically adds effects that are significantly associated with the target. However, after an effect is added to the model, Stepwise can remove any effect already in the model that is not significantly associated with the target. This stepwise process continues until one of the following occurs:

- No other effect in the model meets the Stay Significance Level.
- The Max Steps criterion is met. If you choose the Stepwise selection method, then you can specify a Max Steps to put a limit on the number of steps before the effect selection process stops. The default value is set to the number of effects in the model. If you add interactions via the Interaction Builder, the Max Steps is automatically updated to include these terms.
- An effect added in one step is the only effect deleted in the next step



# Answer 4: Step 19 – Classification Table for Validation Set

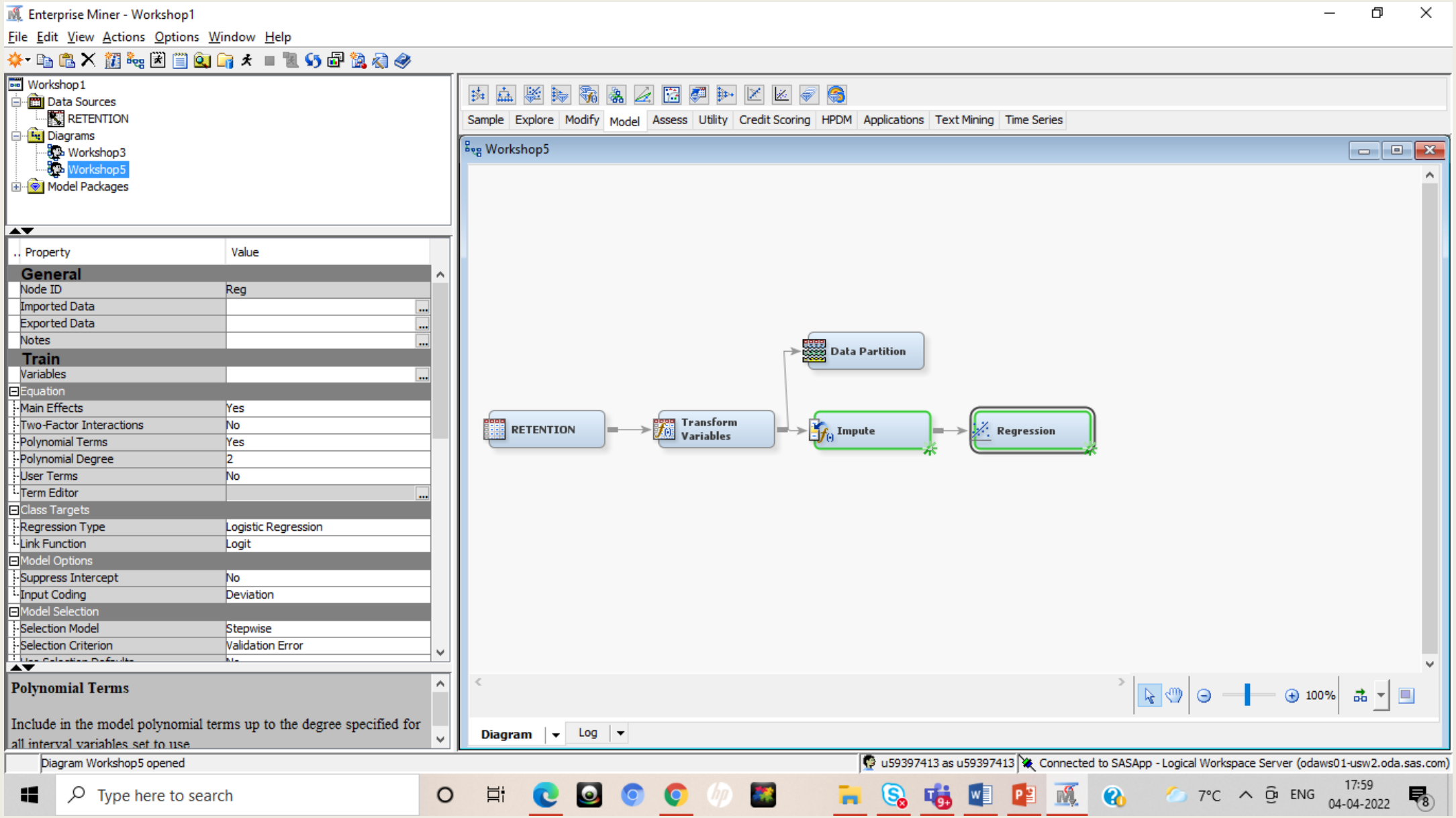
Please find the Classification Table for the Validation Set as follows:

| Misclassification Tree | | | |
|---|---|---|---|
| | Detected as 0 (outcome= 0) | Detected as 1 (outcome = 1) | Total |
| **Truly 0 (target = 0)** | TN= 903 | FP= 87 | FP+TN = 990 |
| **Truly 1 (target = 1)** | FN= 15 | TP= 47 | TP+FN = 62 |
| **Total** | TN+FN= 918 | TP+FP= 134 | |

Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 47/62 = 0.758
**Precision (P)** = TP/(TP+FP) = 47/134 = 0.351
F1 = 2P.R/(P+R) = (2*(0.758*0.351))/(0.758+0.351) = 0.4978

The Precision value is 0.351, which signifies that only 35.1% of the validation dataset is correctly identified as True Positives which is a really low number. The F1 value of 49.78% is low too

# Answer 5: To include Quadratic Polynomial Term in the model



# Answer 6: F1 value after including Quadratic Polynomial Term

Please find the Classification Table for the Validation Set as follows:

| Misclassification Tree | | | |
|---|---|---|---|
| | Detected as 0 (outcome= 0) | Detected as 1 (outcome = 1) | Total |
| **Truly 0 (target = 0)** | TN= 892 | FP= 26 | FP+TN = 918 |
| **Truly 1 (target = 1)** | FN= 75 | TP= 59 | TP+FN = 134 |
| **Total** | TN+FN= 967 | TP+FP= 85 | |

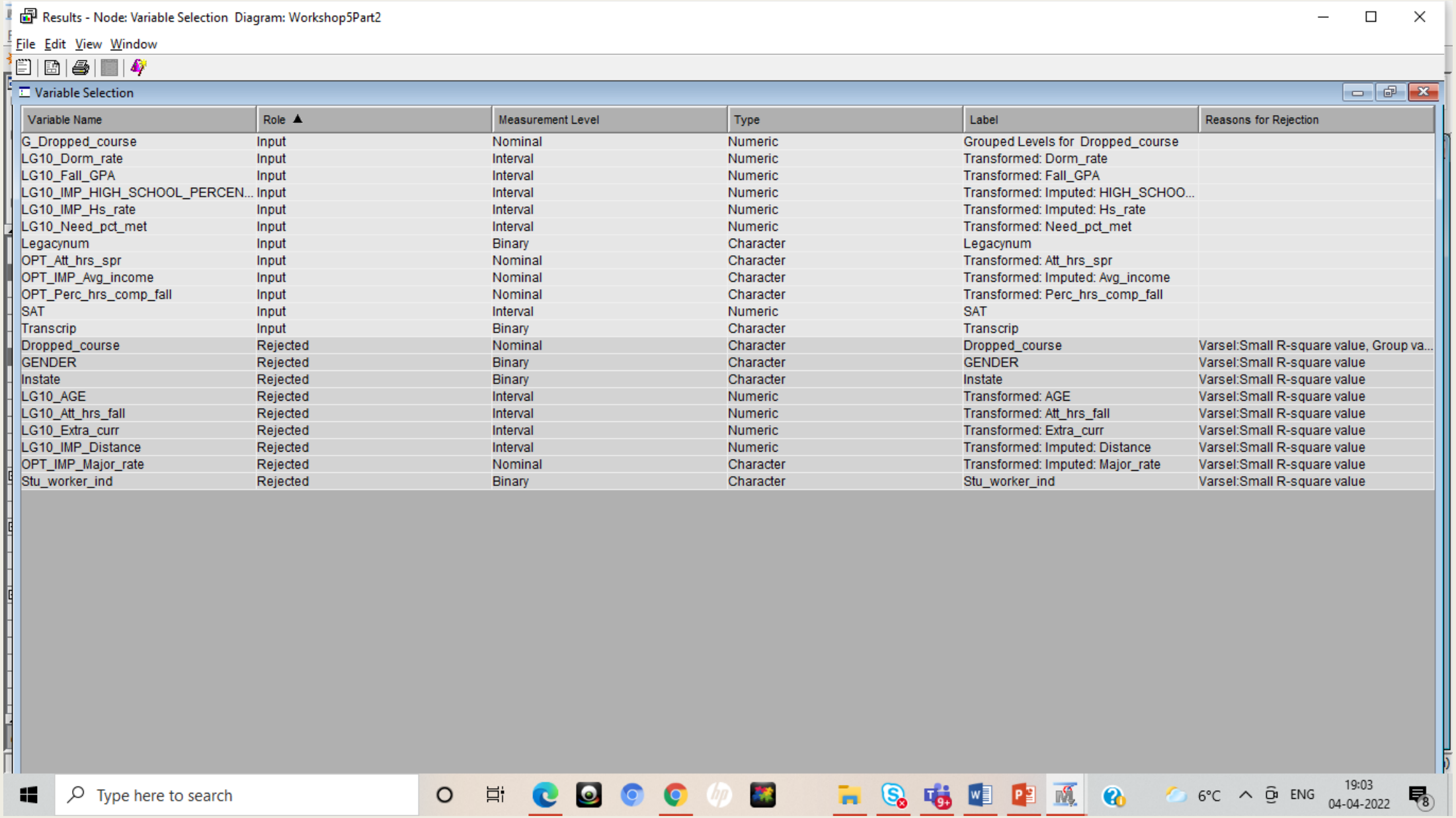
Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 59/134 = 0.440
**Precision (P)** = TP/(TP+FP) = 59/85 = 0.694
F1 = 2P.R/(P+R) = (2*(0.694*0.440))/(0.694+0.440) = 0.538

The Precision value is 0.694, which signifies that only 69.4% of the validation dataset is correctly identified as True Positives which is a better model than the previous model. The F1 score is 53.8% which is higher than 49.78% in the previous model

# PART III: NEURAL NETWORKS

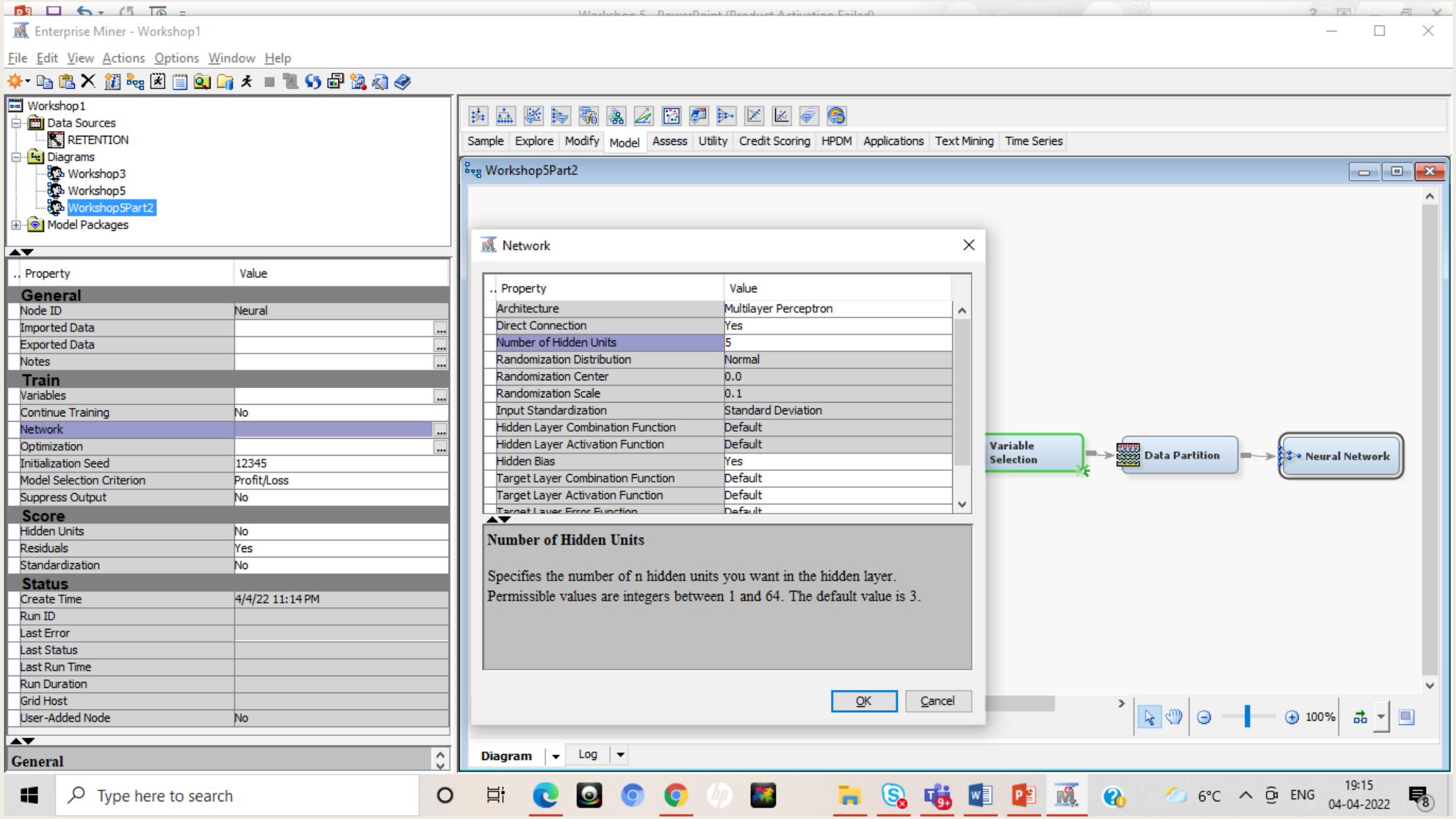
# Answer 7: The number of variables selected in Variable Selection node

Variables with low R square values are rejected. 12 variables are selected




| Variable Name | Role ▲ | Measurement Level | Type | Label | Reasons for Rejection |
|---|---|---|---|---|---|
| G_Dropped_course | Input | Nominal | Numeric | Grouped Levels for Dropped_course | |
| LG10_Dorm_rate | Input | Interval | Numeric | Transformed: Dorm_rate | |
| LG10_Fall_GPA | Input | Interval | Numeric | Transformed: Fall_GPA | |
| LG10_IMP_HIGH_SCHOOL_PERCEN... | Input | Interval | Numeric | Transformed: Imputed: HIGH_SCHOO... | |
| LG10_IMP_Hs_rate | Input | Interval | Numeric | Transformed: Imputed: Hs_rate | |
| LG10_Need_pct_met | Input | Interval | Numeric | Transformed: Need_pct_met | |
| Legacynum | Input | Binary | Character | Legacynum | |
| OPT_Att_hrs_spr | Input | Nominal | Character | Transformed: Att_hrs_spr | |
| OPT_IMP_Avg_income | Input | Nominal | Character | Transformed: Imputed: Avg_income | |
| OPT_Perc_hrs_comp_fall | Input | Nominal | Character | Transformed: Perc_hrs_comp_fall | |
| SAT | Input | Interval | Numeric | SAT | |
| Transcrip | Input | Binary | Character | Transcrip | |
| Dropped_course | Rejected | Nominal | Character | Dropped_course | Varsel:Small R-square value, Group va... |
| GENDER | Rejected | Binary | Character | GENDER | Varsel:Small R-square value |
| Instate | Rejected | Binary | Character | Instate | Varsel:Small R-square value |
| LG10_AGE | Rejected | Interval | Numeric | Transformed: AGE | Varsel:Small R-square value |
| LG10_Att_hrs_fall | Rejected | Interval | Numeric | Transformed: Att_hrs_fall | Varsel:Small R-square value |
| LG10_Extra_curr | Rejected | Interval | Numeric | Transformed: Extra_curr | Varsel:Small R-square value |
| LG10_IMP_Distance | Rejected | Interval | Numeric | Transformed: Imputed: Distance | Varsel:Small R-square value |
| OPT_IMP_Major_rate | Rejected | Nominal | Character | Transformed: Imputed: Major_rate | Varsel:Small R-square value |
| Stu_worker_ind | Rejected | Binary | Character | Stu_worker_ind | Varsel:Small R-square value |

# Answer 8: Neural Network Configuration



# Answer 9: F1 Measure for the Neural Network model

```
Event Classification Table

Data Role=TRAIN Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

   137        1806         27          130


Data Role=VALIDATE Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

   32          456          3          35
```

# Answer 9: F1 Measure for the Neural Network model

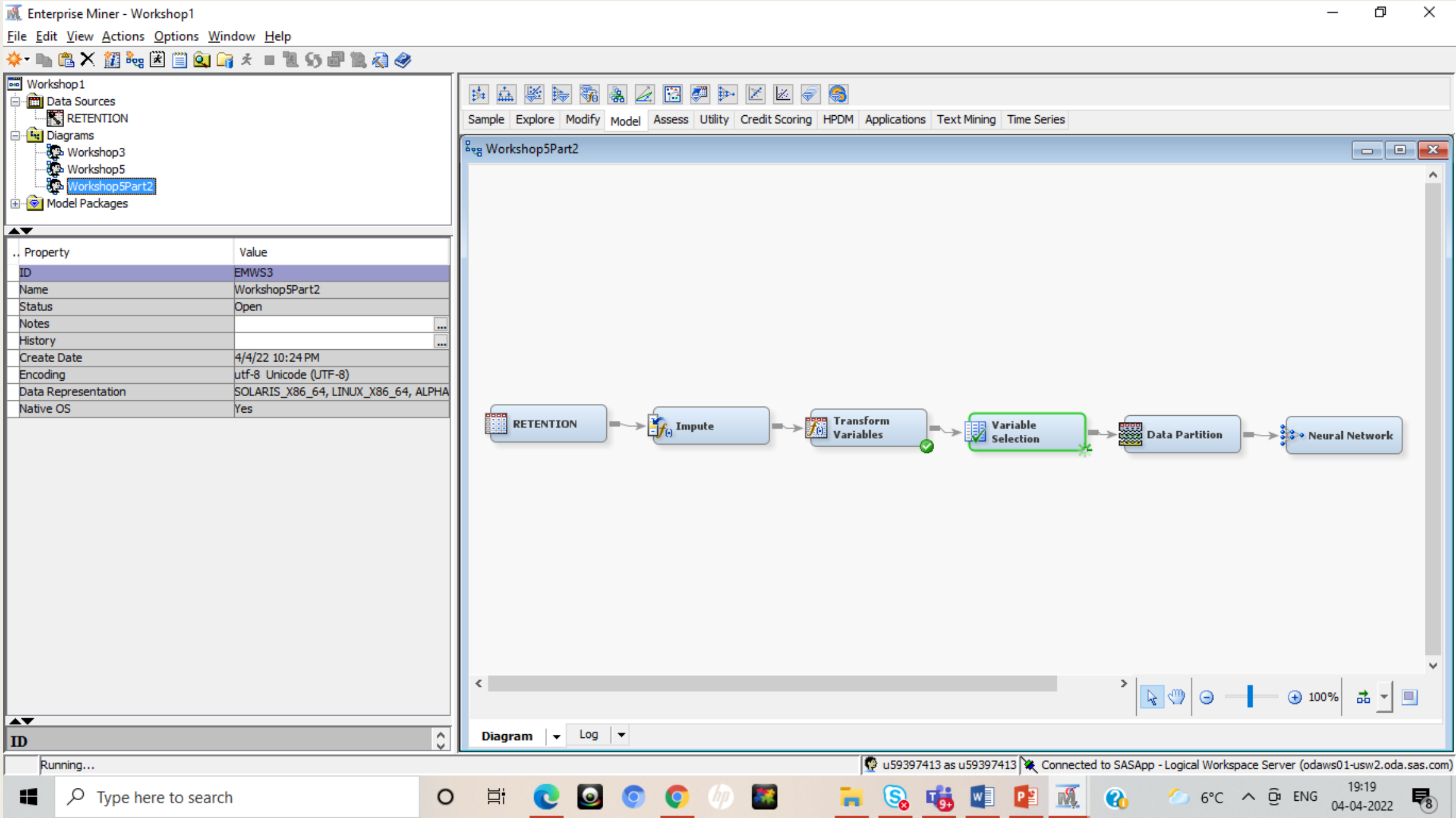
Please find the Classification Table for the Validation Set as follows:

| Misclassification Tree | | | |
|---|---|---|---|
| | Detected as 0 (outcome= 0) | Detected as 1 (outcome = 1) | Total |
| **Truly 0 (target = 0)** | TN= 456 | FP= 3 | FP+TN = 459 |
| **Truly 1 (target = 1)** | FN= 32 | TP= 35 | TP+FN = 67 |
| **Total** | TN+FN= 488 | TP+FP= 38 | |

Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 35/67 = 0.522
**Precision (P)** = TP/(TP+FP) = 35/38 = 0.921
F1 = 2P.R/(P+R) = (2*(0.522*0.921))/(0.522+0.921) = 0.667

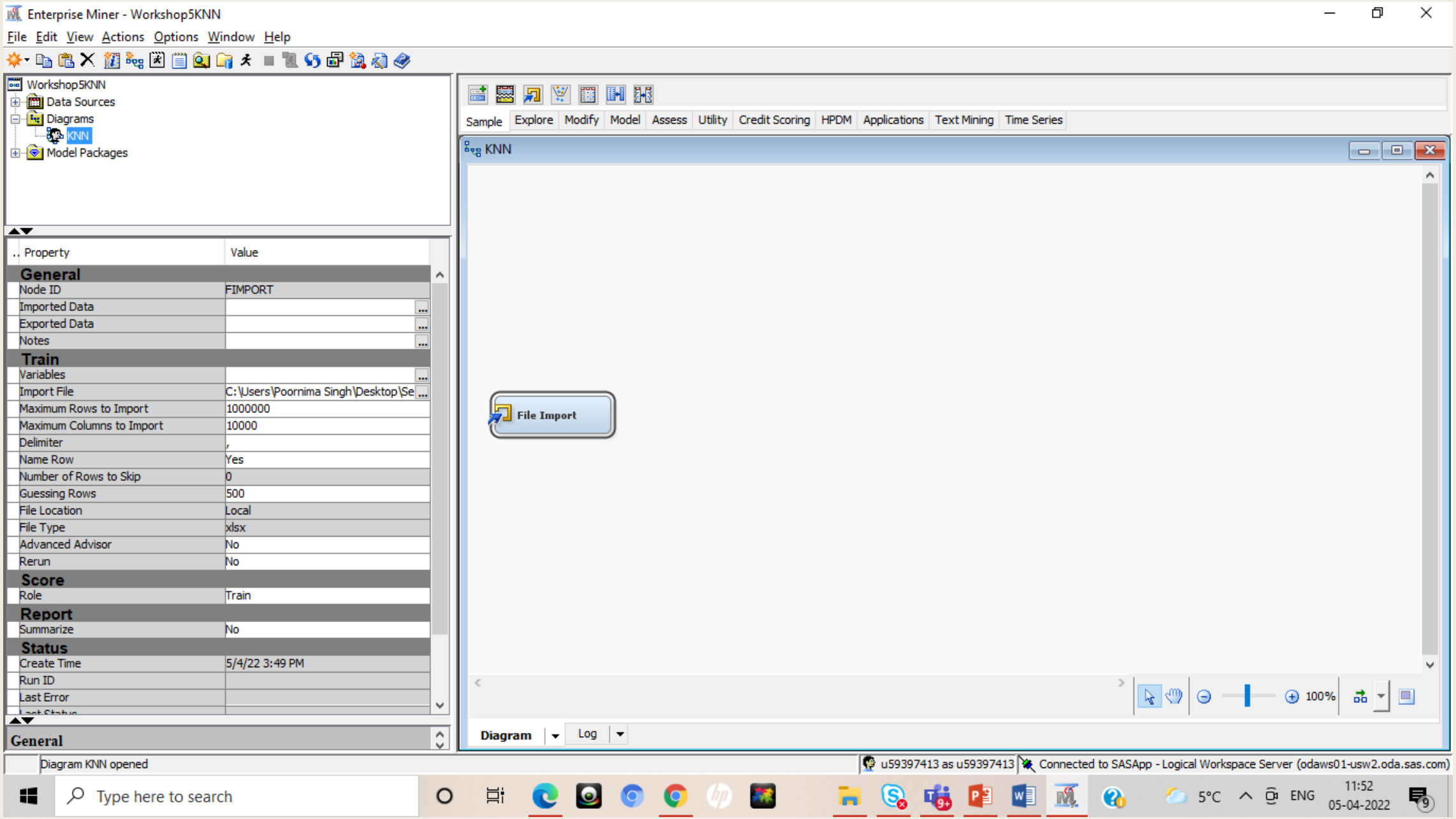
The Precision value is 0.921, which signifies that 92.1% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 66.7% which is higher and better than 53.8% obtained in the regression model

# Answer 10: F1 Measure for the Neural Network model



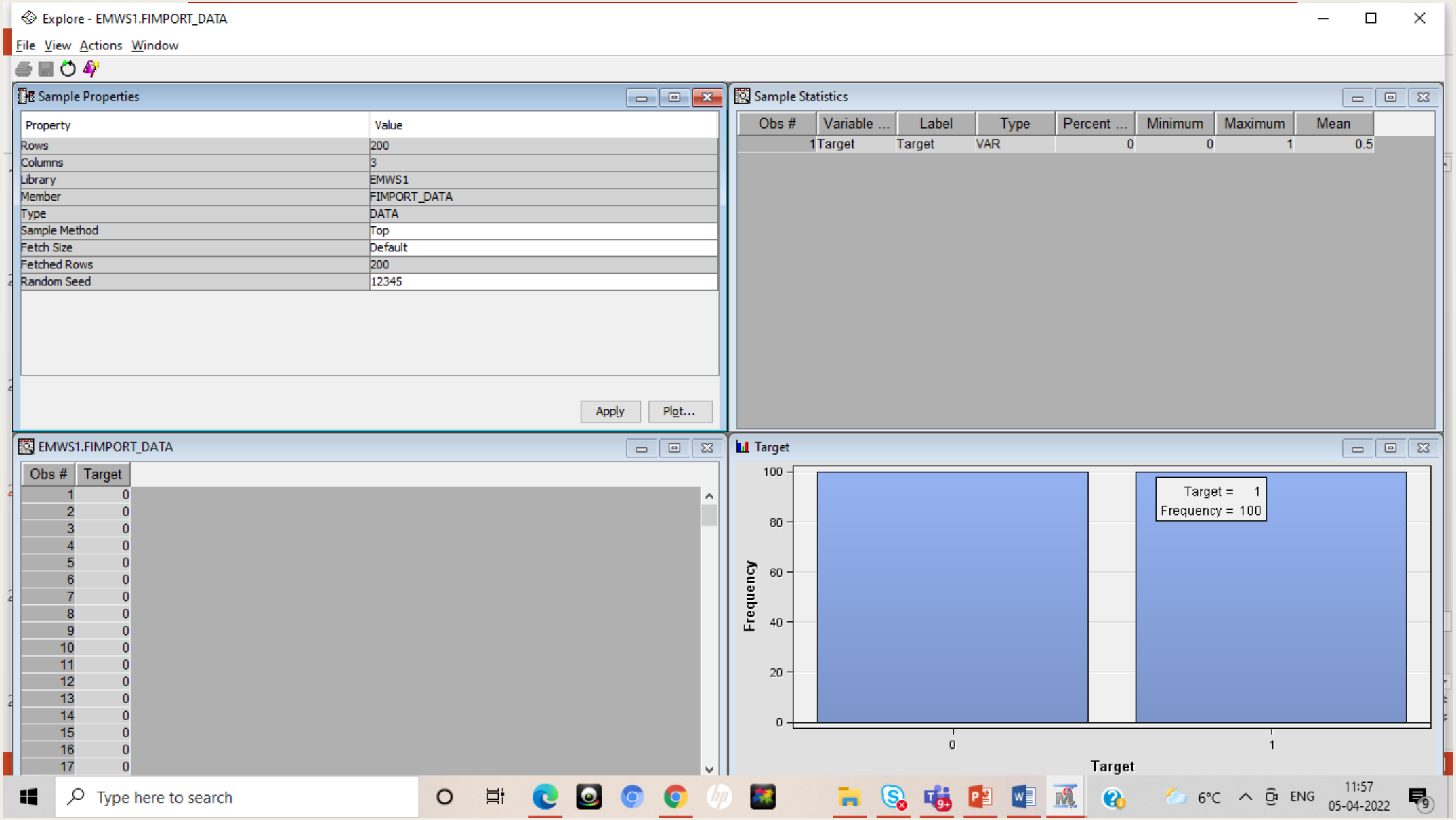
# PART IV: KNN

# Answer 11, 12 and 13: New Project, KNN Diagram and File Import



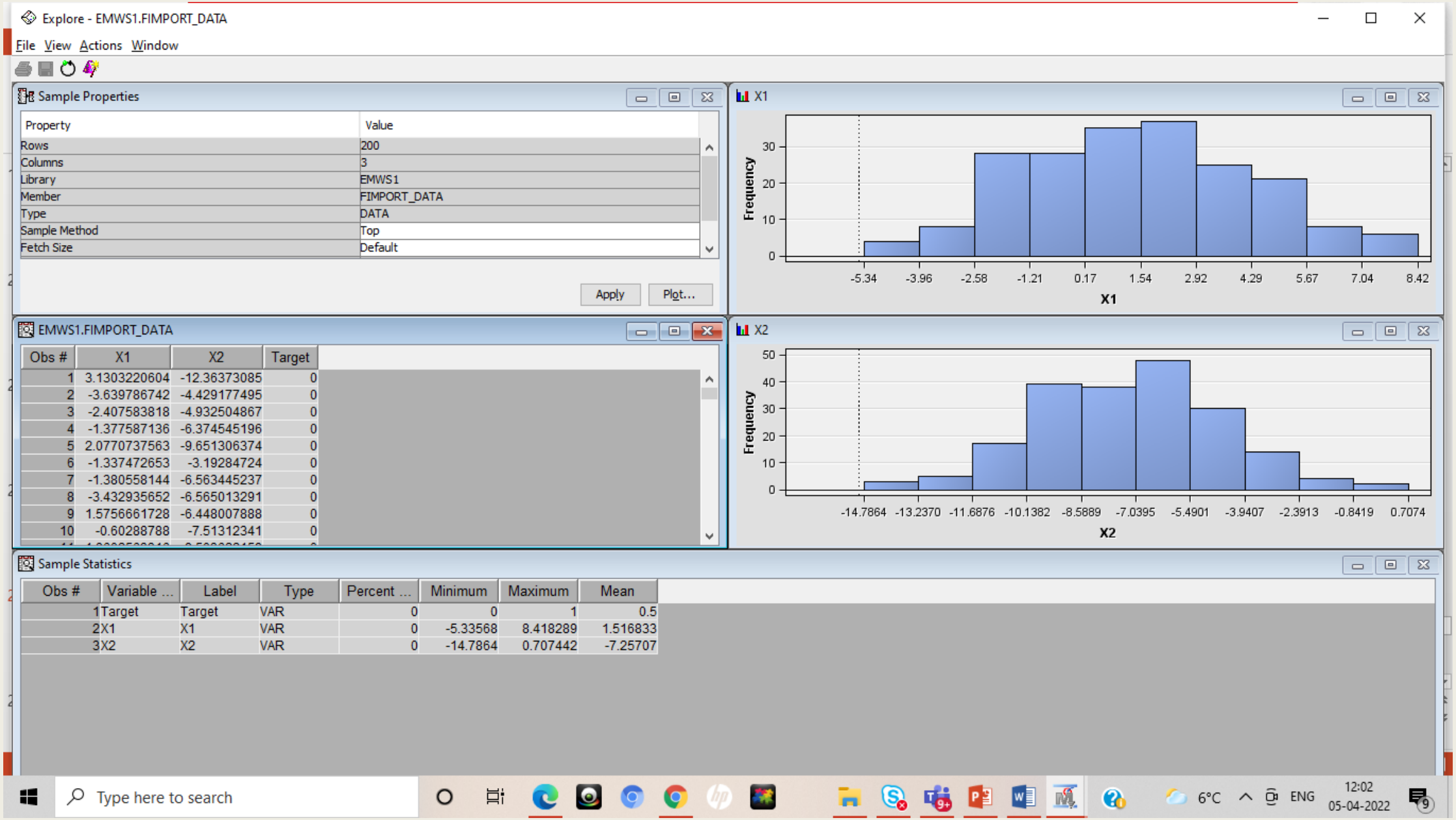
# Answer 14 Part 1: How many rows are imported and how many have Target =1

200 rows were imported out of which 100 have Target = 1



# Answer 14 Part 2: Are the distributions of X1 and X2 normal

X1 and X2 are normally distributed



# Answer 15: How many data samples are in the training set. What percentage has Target = 1?

There are 158 data samples in the training data set, out which 50 percent have Target = 1

Summary Statistics for Class Targets

Data=DATA

| Variable | Numeric Value | Formatted Value | Frequency Count | Percent | Label |
|---|---|---|---|---|---|
| Target | 0 | 0 | 100 | 50 | Target |
| Target | 1 | 1 | 100 | 50 | Target |

Data=TRAIN

| Variable | Numeric Value | Formatted Value | Frequency Count | Percent | Label |
|---|---|---|---|---|---|
| Target | 0 | 0 | 79 | 50 | Target |
| Target | 1 | 1 | 79 | 50 | Target |

Data=VALIDATE

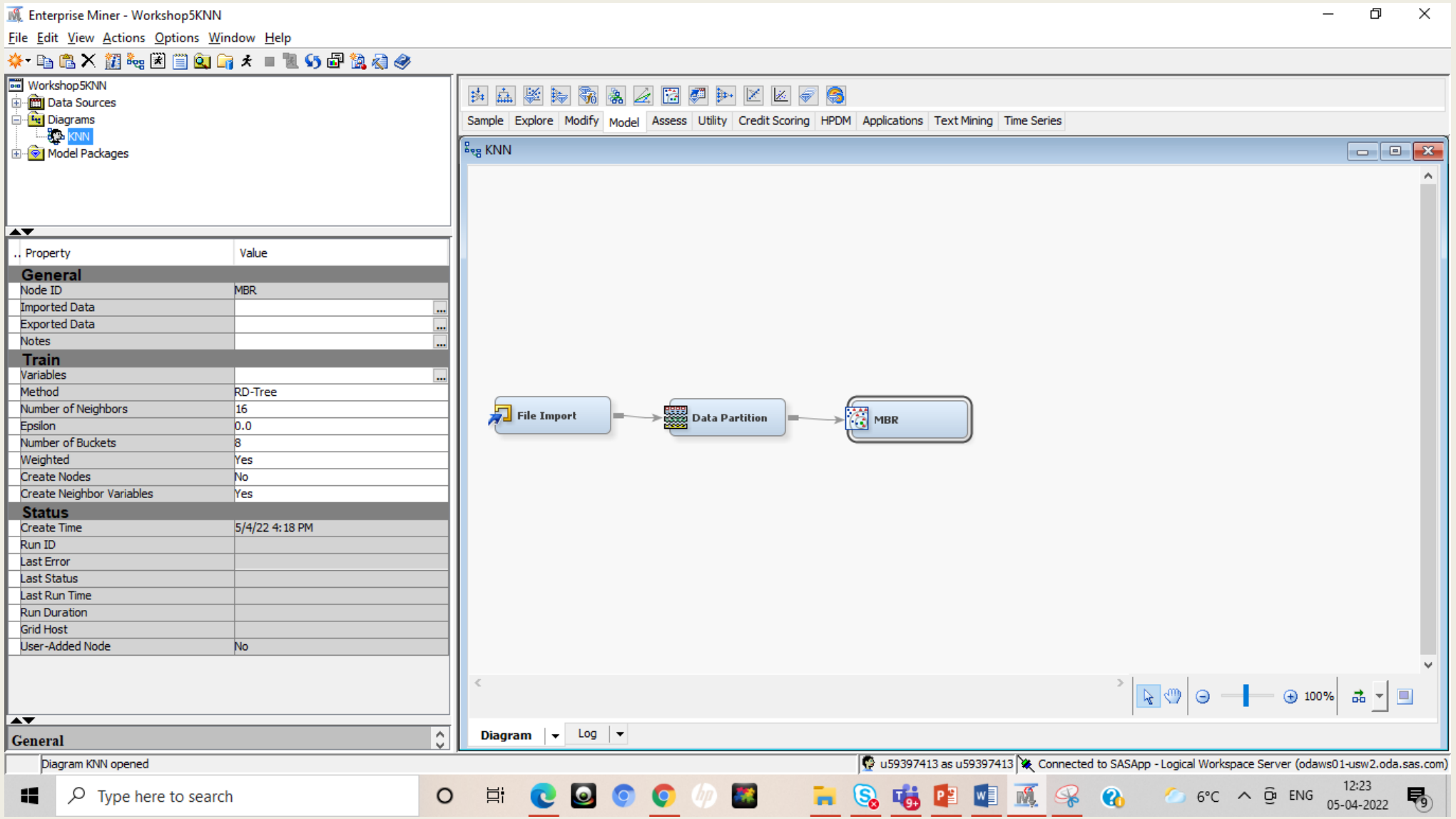
| Variable | Numeric Value | Formatted Value | Frequency Count | Percent | Label |
|---|---|---|---|---|---|
| Target | 0 | 0 | 21 | 50 | Target |
| Target | 1 | 1 | 21 | 50 | Target |

Partition Summary

| Type | Data Set | Number of Observations |
|---|---|---|
| DATA | EMWS1.FIMPORT_train | 200 |
| TRAIN | EMWS1.Part_TRAIN | 158 |
| VALIDATE | EMWS1.Part_VALIDATE | 42 |

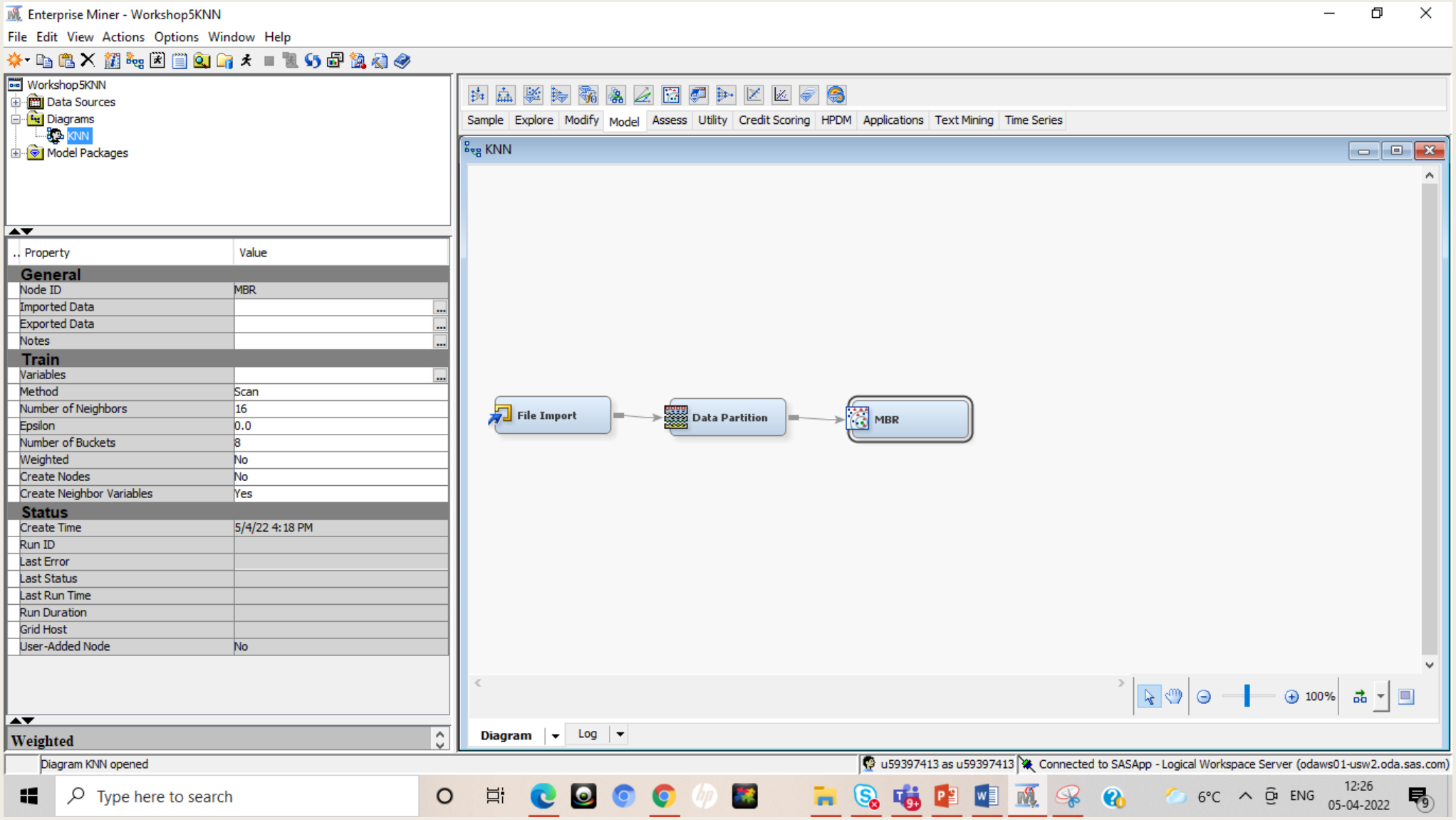
# Answer 18 Part 1: What is the default method under Train?

The default method under Train is RD-TREE



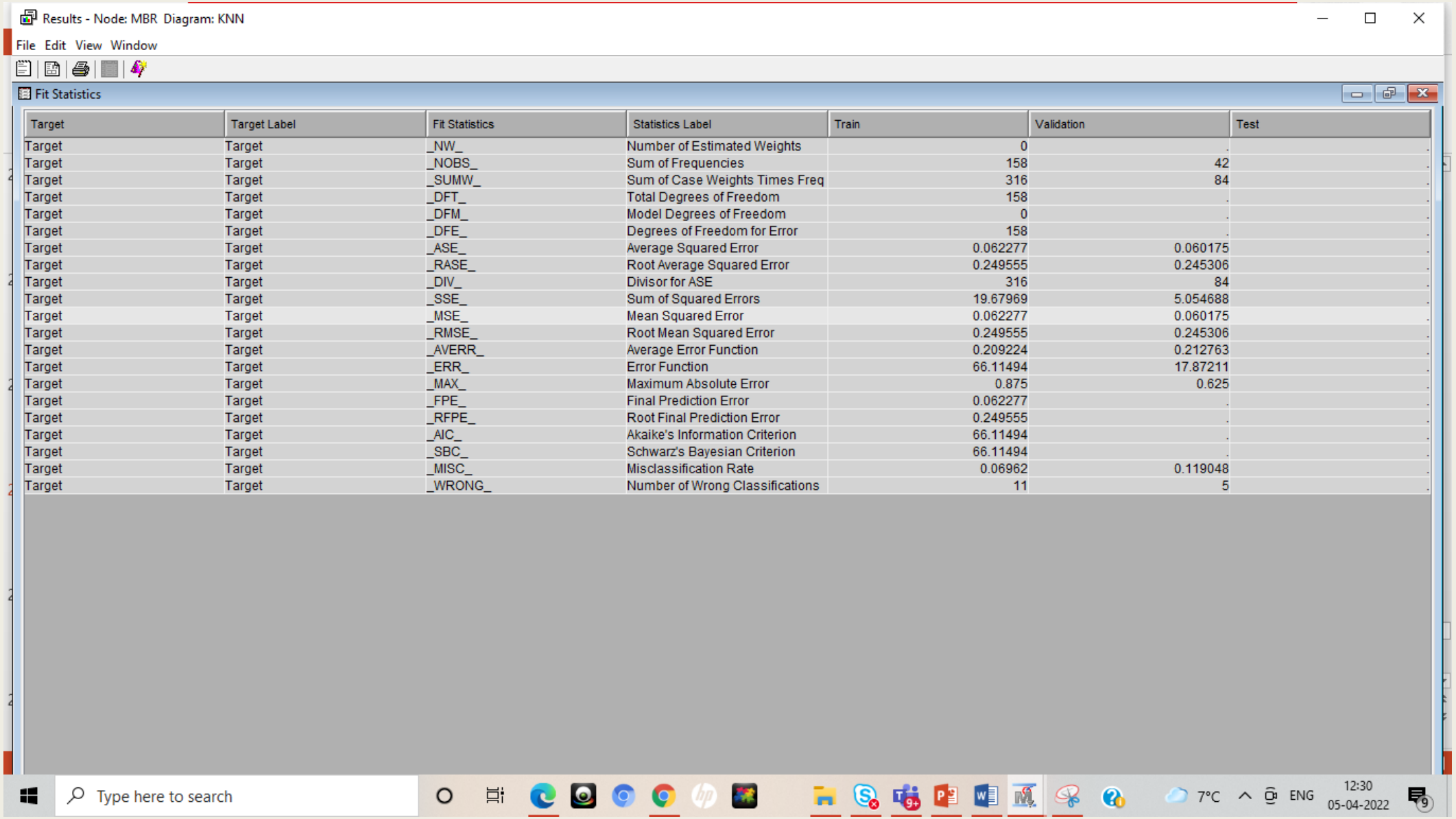
# Answer 18 Part 2: What is the default value for k (number of neighbors)?

The default value of k is 16



# Answer 19 Part 1: What is the MSE for training versus validation data set?

The MSE is as follows:
Train dataset = 0.062277 and Validation dataset = 0.060175




| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Target | Target | _NW_ | Number of Estimated Weights | 0 | . | . |
| Target | Target | _NOBS_ | Sum of Frequencies | 158 | 42 | . |
| Target | Target | _SUMW_ | Sum of Case Weights Times Freq | 316 | 84 | . |
| Target | Target | _DFT_ | Total Degrees of Freedom | 158 | . | . |
| Target | Target | _DFM_ | Model Degrees of Freedom | 0 | . | . |
| Target | Target | _DFE_ | Degrees of Freedom for Error | 158 | . | . |
| Target | Target | _ASE_ | Average Squared Error | 0.062277 | 0.060175 | . |
| Target | Target | _RASE_ | Root Average Squared Error | 0.249555 | 0.245306 | . |
| Target | Target | _DIV_ | Divisor for ASE | 316 | 84 | . |
| Target | Target | _SSE_ | Sum of Squared Errors | 19.67969 | 5.054688 | . |
| Target | Target | _MSE_ | Mean Squared Error | 0.062277 | 0.060175 | . |
| Target | Target | _RMSE_ | Root Mean Squared Error | 0.249555 | 0.245306 | . |
| Target | Target | _AVERR_ | Average Error Function | 0.209224 | 0.212763 | . |
| Target | Target | _ERR_ | Error Function | 66.11494 | 17.87211 | . |
| Target | Target | _MAX_ | Maximum Absolute Error | 0.875 | 0.625 | . |
| Target | Target | _FPE_ | Final Prediction Error | 0.062277 | . | . |
| Target | Target | _RFPE_ | Root Final Prediction Error | 0.249555 | . | . |
| Target | Target | _AIC_ | Akaike's Information Criterion | 66.11494 | . | . |
| Target | Target | _SBC_ | Schwarz's Bayesian Criterion | 66.11494 | . | . |
| Target | Target | _MISC_ | Misclassification Rate | 0.06962 | 0.119048 | . |
| Target | Target | _WRONG_ | Number of Wrong Classifications | 11 | 5 | . |

# Answer 19 Part 2: Find the number of FN, FP, etc. for the validation set and paste here

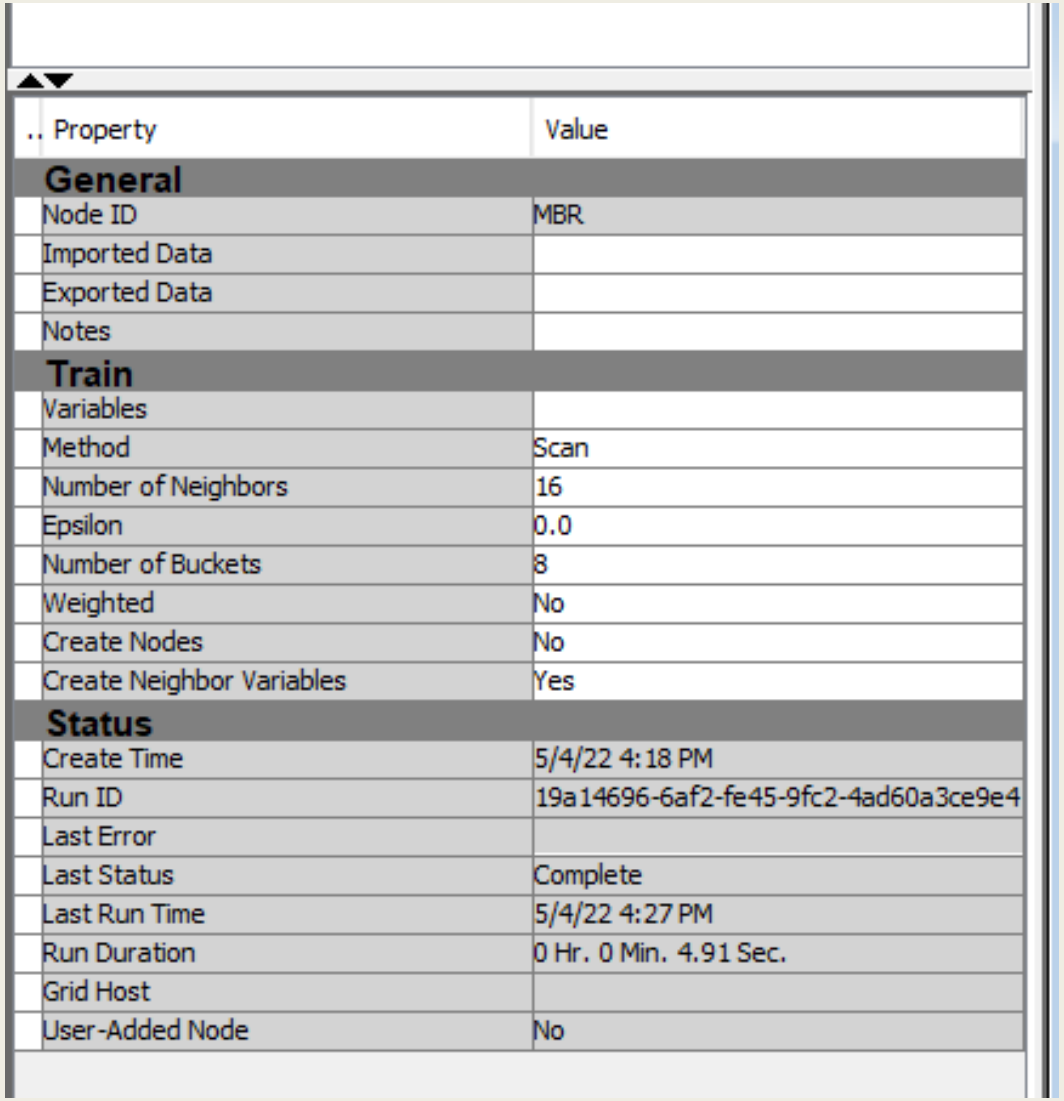
```
Event Classification Table

Data Role=TRAIN Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

    8          76          3          71


Data Role=VALIDATE Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

    3          19          2          18
```

Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 18/(18+19) = 0.486
**Precision (P)** = TP/(TP+FP) = 18/(18+2) = 0.90
F1 = 2P.R/(P+R) = (2*(0.486*0.90))/(0.486+0.90) = 0.631

The Precision value is 0.90, which signifies that 90.0% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 63.1%.

# Answer 19 Part 3: Run Duration of the model

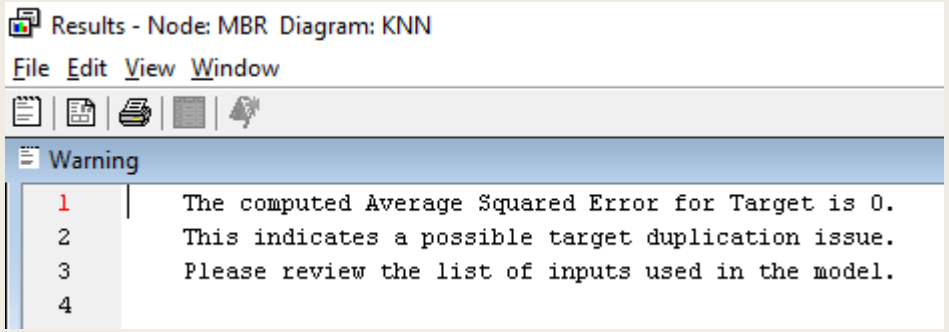
The run duration of the model is : 0 Hr. 0 Min. 4.91 Sec.



| .. Property | Value |
|---|---|
| **General** | |
| Node ID | MBR |
| Imported Data | |
| Exported Data | |
| Notes | |
| **Train** | |
| Variables | |
| Method | Scan |
| Number of Neighbors | 16 |
| Epsilon | 0.0 |
| Number of Buckets | 8 |
| Weighted | No |
| Create Nodes | No |
| Create Neighbor Variables | Yes |
| **Status** | |
| Create Time | 5/4/22 4:18 PM |
| Run ID | 19a14696-6af2-fe45-9fc2-4ad60a3ce9e4 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 5/4/22 4:27 PM |
| Run Duration | 0 Hr. 0 Min. 4.91 Sec. |
| Grid Host | |
| User-Added Node | No |

# Answer 20 Part 1: What is the warning and what is the MSE value this time?

The MSE is as follows:
Train dataset = 0 and Validation dataset = 0.119048




The warning is as shown on the left

Results - Node: MBR Diagram: KNN

File Edit View Window

Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Target | Target | _NW_ | Number of Estimated Weights | 0 | . | . |
| Target | Target | _NOBS_ | Sum of Frequencies | 158 | 42 | . |
| Target | Target | _SUMW_ | Sum of Case Weights Times Freq | 316 | 84 | . |
| Target | Target | _DFT_ | Total Degrees of Freedom | 158 | . | . |
| Target | Target | _DFM_ | Model Degrees of Freedom | 0 | . | . |
| Target | Target | _DFE_ | Degrees of Freedom for Error | 158 | . | . |
| Target | Target | _ASE_ | Average Squared Error | 0 | 0.119048 | . |
| Target | Target | _RASE_ | Root Average Squared Error | 0 | 0.345033 | . |
| Target | Target | _DIV_ | Divisor for ASE | 316 | 84 | . |
| Target | Target | _SSE_ | Sum of Squared Errors | 0 | 10 | . |
| Target | Target | _MSE_ | Mean Squared Error | 0 | 0.119048 | . |
| Target | Target | _RMSE_ | Root Mean Squared Error | 0 | 0.345033 | . |
| Target | Target | _AVERR_ | Average Error Function | 0 | 0.822352 | . |
| Target | Target | _ERR_ | Error Function | 0 | 69.07755 | . |
| Target | Target | _MAX_ | Maximum Absolute Error | 0 | 1 | . |
| Target | Target | _FPE_ | Final Prediction Error | 0 | . | . |
| Target | Target | _RFPE_ | Root Final Prediction Error | 0 | . | . |
| Target | Target | _AIC_ | Akaike's Information Criterion | . | . | . |
| Target | Target | _SBC_ | Schwarz's Bayesian Criterion | . | . | . |
| Target | Target | _MISC_ | Misclassification Rate | 0 | 0.119048 | . |
| Target | Target | _WRONG_ | Number of Wrong Classifications | 0 | 5 | . |

# Answer 20 Part 2: What is the number of FN, FP, etc. for validation set here. How did the results change?

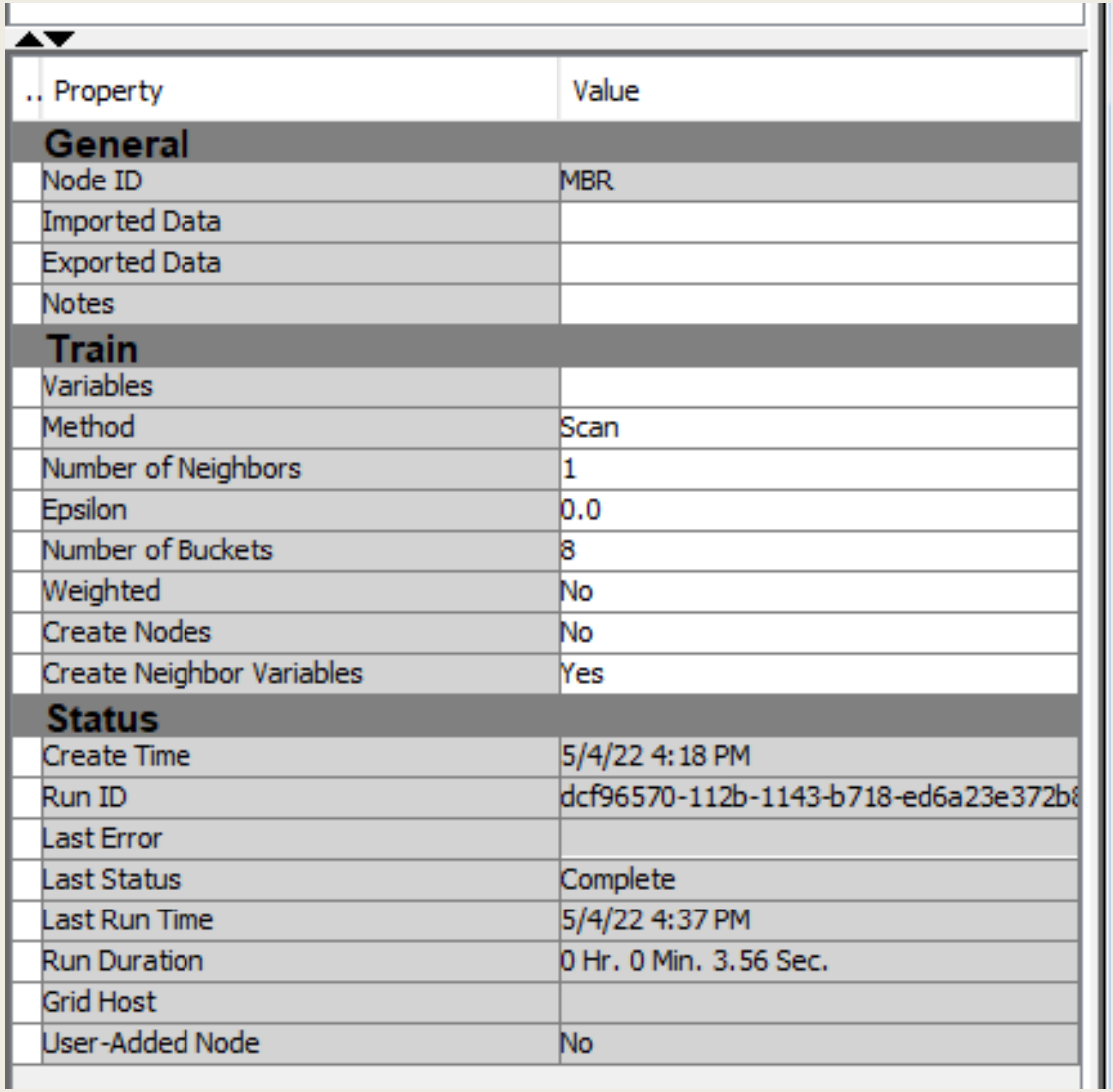
```
Event Classification Table

Data Role=TRAIN Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

    .          79           .          79


Data Role=VALIDATE Target=Target Target Label=Target

  False       True        False       True
Negative    Negative    Positive    Positive

    1          17           4          20
```

Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 17/(17+1) = 0.944
**Precision (P)** = TP/(TP+FP) = 17/(17+4) = 0.8095
F1 = 2P.R/(P+R) = (2*(0.955*0.8095))/(0.8095+0.955) = 0.8664

The Precision value is 0.80, which signifies that 80.95% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 86.64% which is the highest as compared to all the models.

# Answer 20 Part 3: Run Duration of the model

The run duration of the model is : 0 Hr. 0 Min. 3.56 Sec.



| Property | Value |
|---|---|
| **General** | |
| Node ID | MBR |
| Imported Data | |
| Exported Data | |
| Notes | |
| **Train** | |
| Variables | |
| Method | Scan |
| Number of Neighbors | 1 |
| Epsilon | 0.0 |
| Number of Buckets | 8 |
| Weighted | No |
| Create Nodes | No |
| Create Neighbor Variables | Yes |
| **Status** | |
| Create Time | 5/4/22 4:18 PM |
| Run ID | dcf96570-112b-1143-b718-ed6a23e372b8 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 5/4/22 4:37 PM |
| Run Duration | 0 Hr. 0 Min. 3.56 Sec. |
| Grid Host | |
| User-Added Node | No |

# Answer 21: What are Eigen Values and What percentage of the energy is explained by the first principal component?

The DMNEURL Procedure

Eigenvalues of Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 1.64015952 | 1.28031904 | 0.8201 | 0.8201 |
| 2 | 0.35984048 | | 0.1799 | 1.0000 |

```
53
54   *-------------------------------------------------------------*
55   Summary of Exported Principal Components
56   *-------------------------------------------------------------*
57
58   Remark:The number of inputs is used as the maximum number of principal components
59
60   Total number of input variables: 2
61   Maximum number cutoff of principal components: 2
62   Cumulative proportional eigenvalue cutoff: 0.99
63   Proportional eigenvalue increment cutoff: 0.001
64   Number of the selected principal components: 2
65   Total variation explained by the selected principal components: 1
66
```

# Answer 22: Event Classification Table after Principal Component and k=16

```
117
118   Event Classification Table
119
120   Data Role=TRAIN Target=Target Target Label=Target
121
122     False       True        False       True
123   Negative    Negative    Positive    Positive
124
125       8          75           4          71
126
127
128   Data Role=VALIDATE Target=Target Target Label=Target
129
130     False       True        False       True
131   Negative    Negative    Positive    Positive
132
133       1          20           1          20
134
```

Please find the Calculations:
**Recall (R)** = TP/(TP+FN) = 20/(20+1) = 0.952
**Precision (P)** = TP/(TP+FP) = 20/(20+1) = 0.952
F1 = 2P.R/(P+R) = (2*(0.952*0.952))/(0.952+0.952) = 0.952

The Precision value is 0.80, which signifies that 95.20% of the validation dataset is correctly identified as True Positives which is a better model than the Regression model. The F1 score is 95.20% which is better than the previous model.

# Answer 23: KNN Diagram and the improvements in results

# GROUP WORK DECLARATION

We, **Group 5 (Anand Mohan Thankur, Josh Shaji, Poonam Bhaliyan, Prateek Ramjanam Singh, and Poornima Singh)** declare that the attached assignment is our own work in accordance with the Seneca Academic Policy. We have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. We have not distributed our work to other students.

| | Name | Task(s) |
|---|---|---|
| 1 | Anand Mohan Thakur (149200206) | Consolidated the Workshop together on MS Teams |
| 2 | Josh Shaji (133557215) | Consolidated the Workshop together on MS Teams |
| 3 | Poonam Bhaliyan (121114219) | Consolidated the Workshop together on MS Teams |
| 4 | Prateek Ramjanam Singh (124483215) | Consolidated the Workshop together on MS Teams |
| 5 | Poornima Singh (125638213) | Consolidated the Workshop together on MS Teams |

# THANK YOU