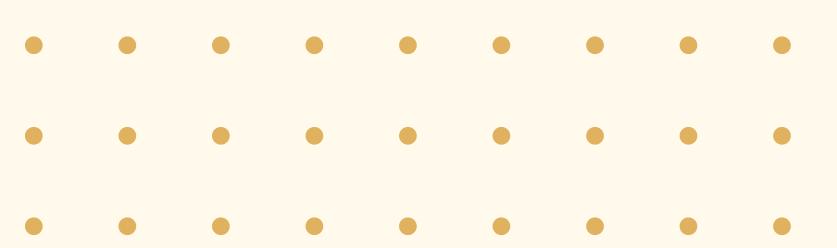


# **WIE3007 DATA MINING AND WAREHOUSING**

**RECORDED PRESENTATION**





# Group Members



Au Wan Ying  
U2005373/1



Chuah Ann Joe  
U2005355/1



Lee Xiao Yu  
U2005405/1



Ruo Jun Wang  
S2011618



Tiew Ker Xin  
U2005253/1



# Business Objectives

O1

To investigate the factors that contribute to the academic dropouts.

O2

To optimise the curriculum performance based on the identified factors.

O3

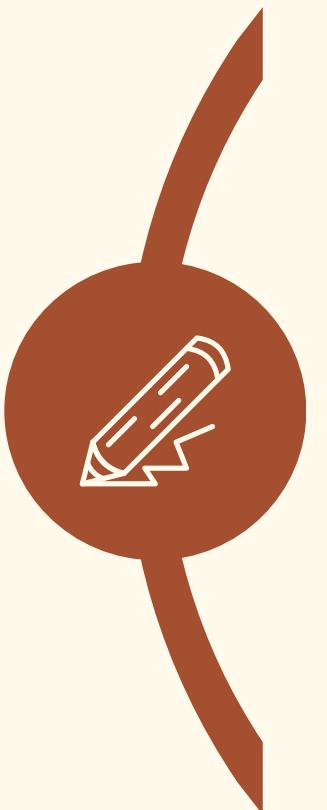
To foster a collaborative learning environment based on the factors that contribute to the academic success

# Introduction

Our dataset presents a project aimed at reducing dropout rates in higher education by identifying at-risk students early on.

The dataset, comprising 4424 records and 36 features , is used for a three-category classification task (dropout, enrolment, and postgraduate).

It encompasses data from various undergraduate fields like agriculture, design, nursing, journalism, and management, gathered from higher education institutions.



# Sample

# Raw Data

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification
0	1	17	5	171	1	1	122.0	1	19	
1	1	15	1	9254	1	1	160.0	1	1	
2	1	1	5	9070	1	1	122.0	1	37	
3	1	17	2	9773	1	1	122.0	1	38	
4	2	39	1	8014	0	1	100.0	1	37	
...	...	...	...	...	...	...	...	...	...	...
<b>4419</b>	1	1	6	9773	1	1	125.0	1	1	
<b>4420</b>	1	1	2	9773	1	1	120.0	105	1	
<b>4421</b>	1	1	1	9500	1	1	154.0	1	37	
<b>4422</b>	1	1	1	9147	1	1	180.0	1	37	
<b>4423</b>	1	10	1	9773	1	1	152.0	22	38	

4424 rows × 37 columns



# Random Sampling



	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification
1255	4	39	1	9130	1	1	133.1	1	3	
3458	1	17	1	9238	1	1	125.0	1	4	
3390	1	17	1	9853	1	1	133.0	1	38	
1497	1	17	2	9670	1	1	110.0	1	1	
1536	1	39	1	9500	1	1	130.0	1	37	
***	***	***	***	***	***	***	***	***	***	***
4235	1	1	3	9238	1	1	133.1	1	1	
979	1	18	1	9003	1	1	133.1	1	1	
283	1	1	1	9070	1	1	126.0	1	19	
1298	1	1	1	9991	0	1	145.0	1	2	
898	1	1	1	9254	1	1	112.0	1	38	

1327 rows × 37 columns



# Stratified Random Sampling



Organizing data into distinct groups based on specified column categories, followed by random sampling within each of these groups.

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	Curricular units 2nd sem (credited)	Curric units (enro)
<b>Marital status</b>												
1	3290	1	1	1	9238	1	1	125.0	1	1	1	...
	3416	1	1	1	9773	1	1	143.0	1	3	3	...
	2522	1	17	6	9500	1	1	131.0	1	1	19	...
	4135	1	1	1	8014	0	1	115.0	1	38	38	...
	981	2	39	1	9991	0	1	100.0	1	19	19	...
2	2006	2	39	1	8014	0	1	146.0	1	37	37	...
	4034	2	39	1	8014	0	19	133.1	1	34	34	...
	1739	2	39	1	9070	1	19	133.1	1	37	37	...
	688	3	39	1	9991	0	1	120.0	1	37	1	...
3	166	3	39	1	9003	1	1	170.0	1	1	37	...
	1428	3	17	1	9238	1	1	138.0	1	37	37	...
	428	3	1	2	9085	1	1	135.0	1	19	1	...
	409	4	39	1	9085	1	1	133.1	1	34	34	...
4	939	4	39	1	8014	0	19	133.1	1	37	37	...
	2999	4	39	1	8014	0	1	138.0	1	19	19	...
	2794	4	39	1	9991	0	1	130.0	1	37	37	...
	4374	5	7	1	9500	1	40	150.0	1	37	37	...
5	3943	5	39	1	9147	1	1	180.0	1	1	1	...
	3721	5	7	1	9070	1	3	140.0	1	37	37	...
	2919	5	39	1	9003	1	19	133.1	1	19	1	...
	2914	6	39	2	9147	1	1	133.1	1	37	37	...
6	2915	6	39	1	8014	0	1	133.1	1	37	37	...
	1180	6	39	1	9500	1	1	133.1	1	37	37	...
	1572	6	39	1	9991	0	12	130.0	1	37	37	0

24 rows x 37 columns



# Weighted Random Sampling

Conducting random sampling where each sample is selected based on predefined weights, which determine the likelihood of each sample's inclusion.

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nationality	Mother's qualification	Father's qualification	...	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	(e)
1681	1	1	1	9773	1	1	117.0	1	1	19	...	0	6	
4212	1	17	2	9773	1	1	125.0	1	1	1	...	0	6	
3261	1	1	3	9147	1	1	115.0	1	1	38	...	0	5	
2667	2	39	1	9991	0	19	133.1	1	19	19	...	0	5	
706	2	39	1	8014	0	1	133.1	1	19	1	...	3	7	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
872	1	42	1	9991	0	1	150.0	1	19	19	...	6	11	
1516	1	7	1	9500	1	5	140.0	1	19	1	...	0	8	
3799	2	39	1	9085	1	12	133.1	1	1	39	...	0	6	
964	1	1	1	9773	1	1	121.0	1	1	3	...	0	6	
321	1	42	1	9500	1	1	100.0	1	37	37	...	0	8	

2212 rows × 37 columns



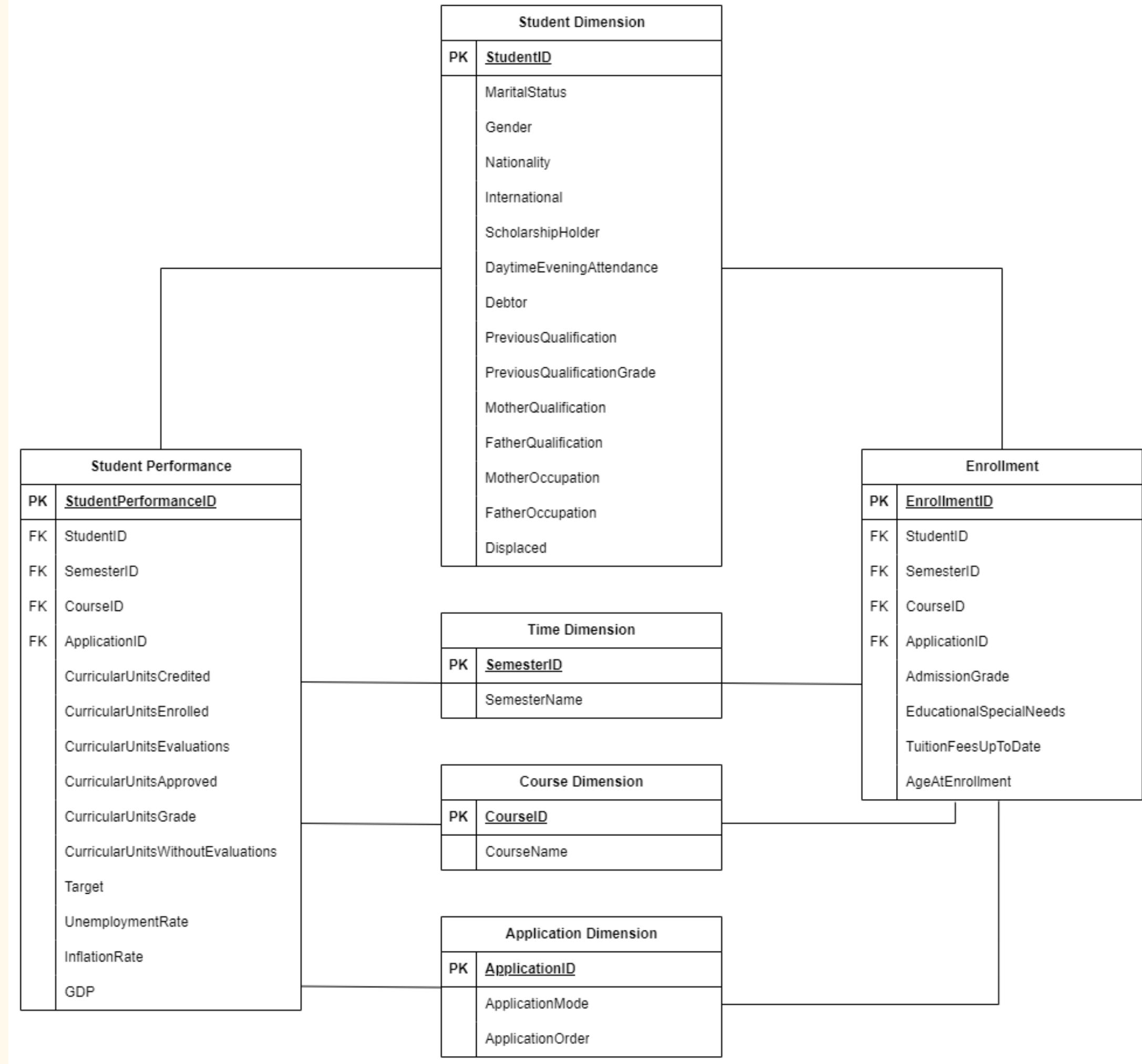
# **Featuretools And Star Schema**

# Featuretools

- Read dataset from csv file
- Define entities
- Create entities set
- Add dataframe to entity set
- Establish relationship
- Perform deep feature synthesis



# Star Schema



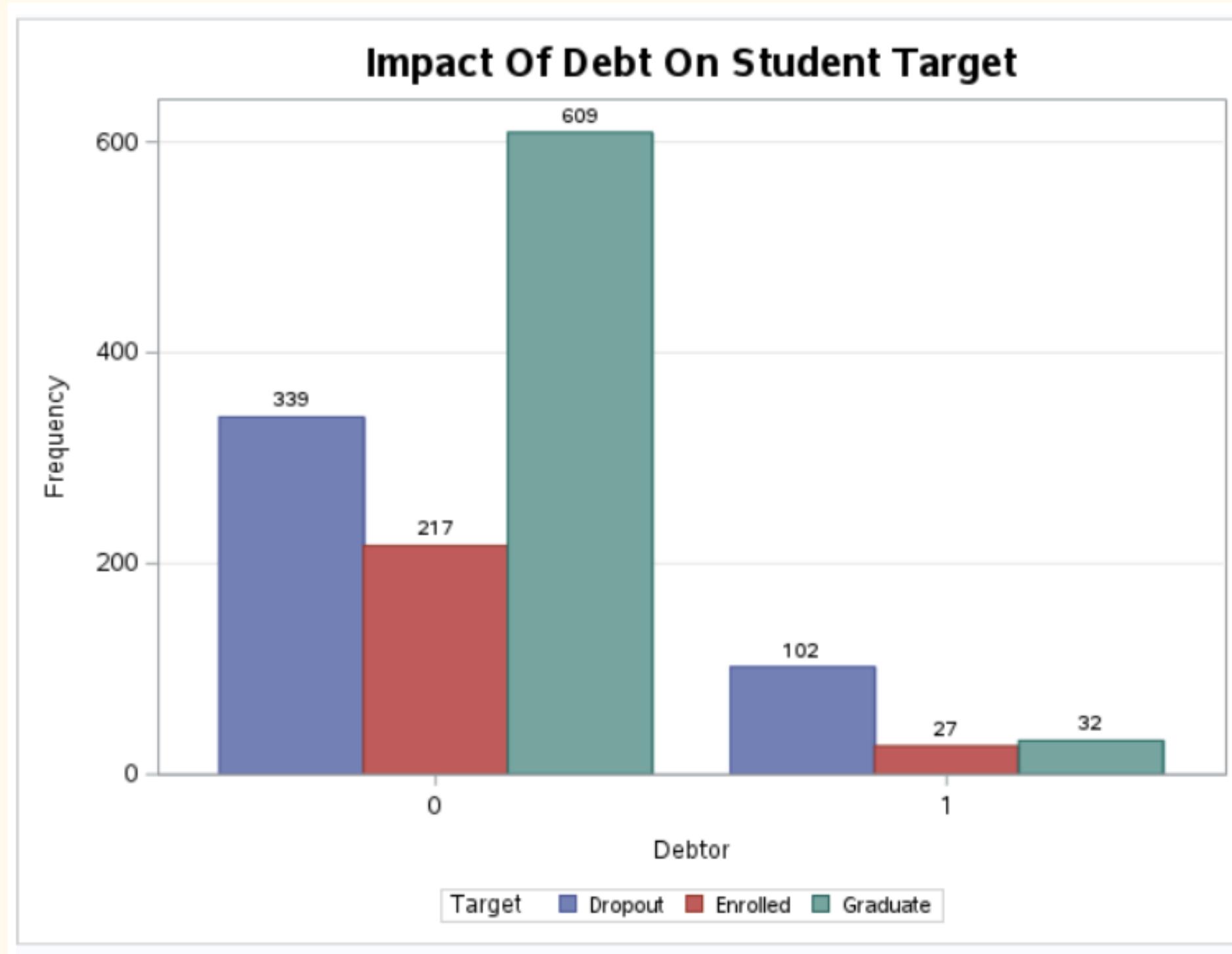
# Explore

# Extract Data From CSV File Using Talend Open Studio



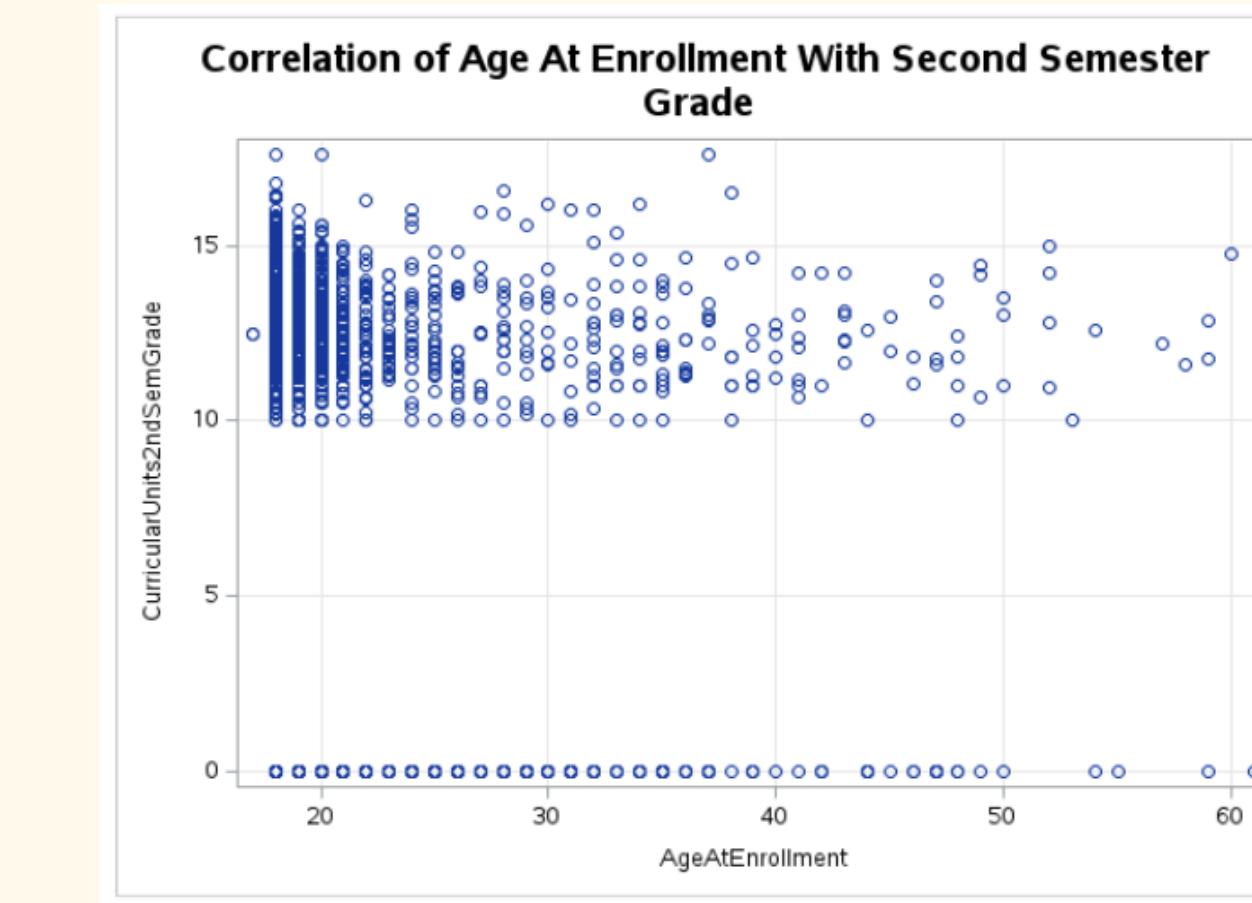
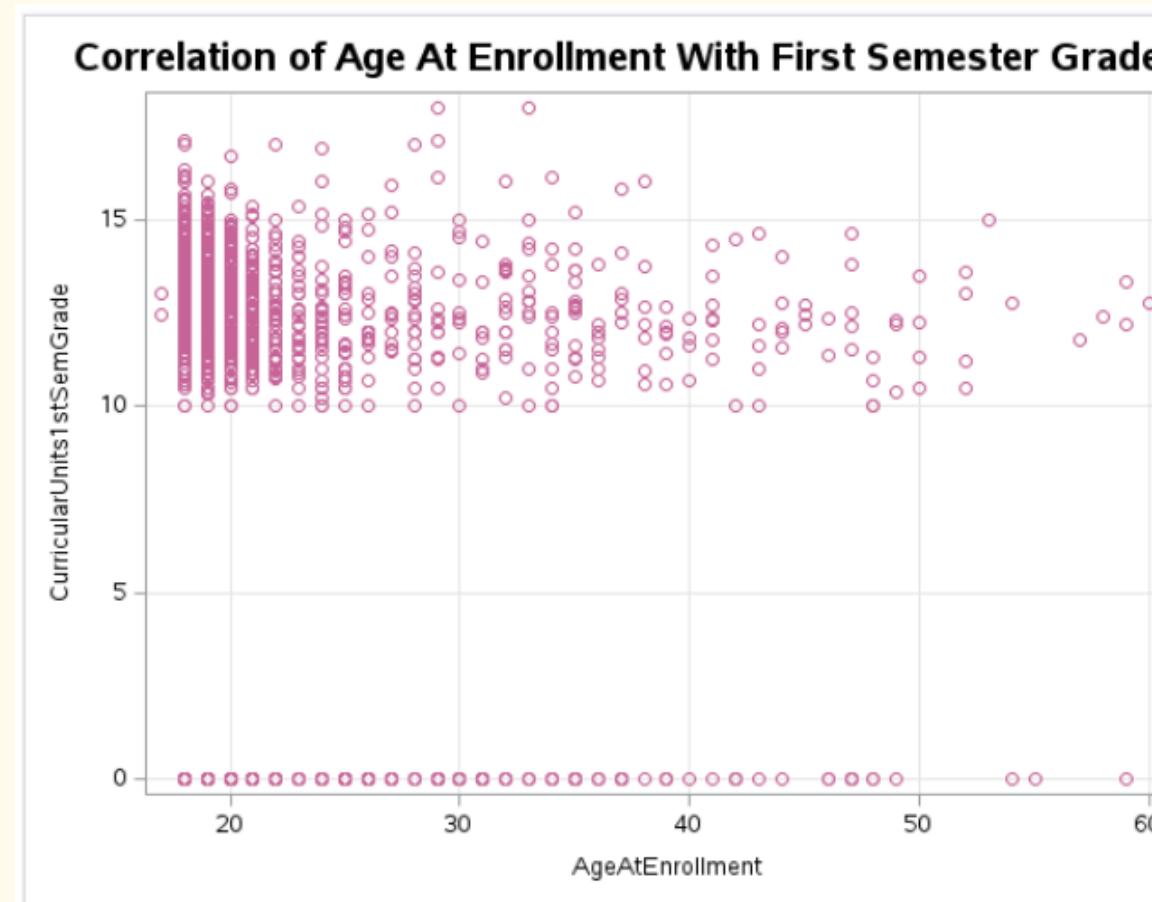
# Exploratory Data Analysis

## Impact Of Debt On Student Target



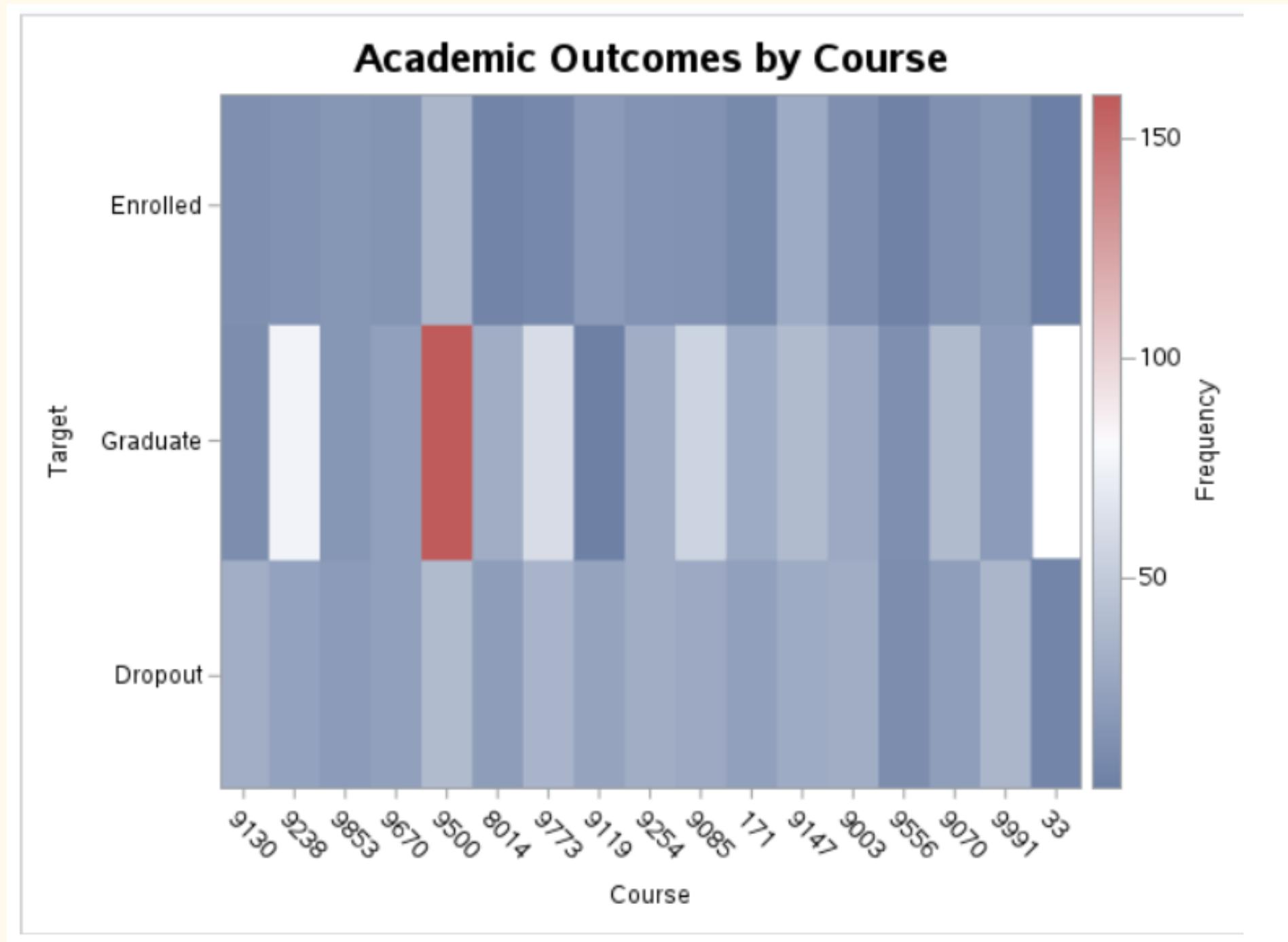
# Exploratory Data Analysis

## Correlation of Age At Enrollment With First & Second Semester Grade



# Exploratory Data Analysis

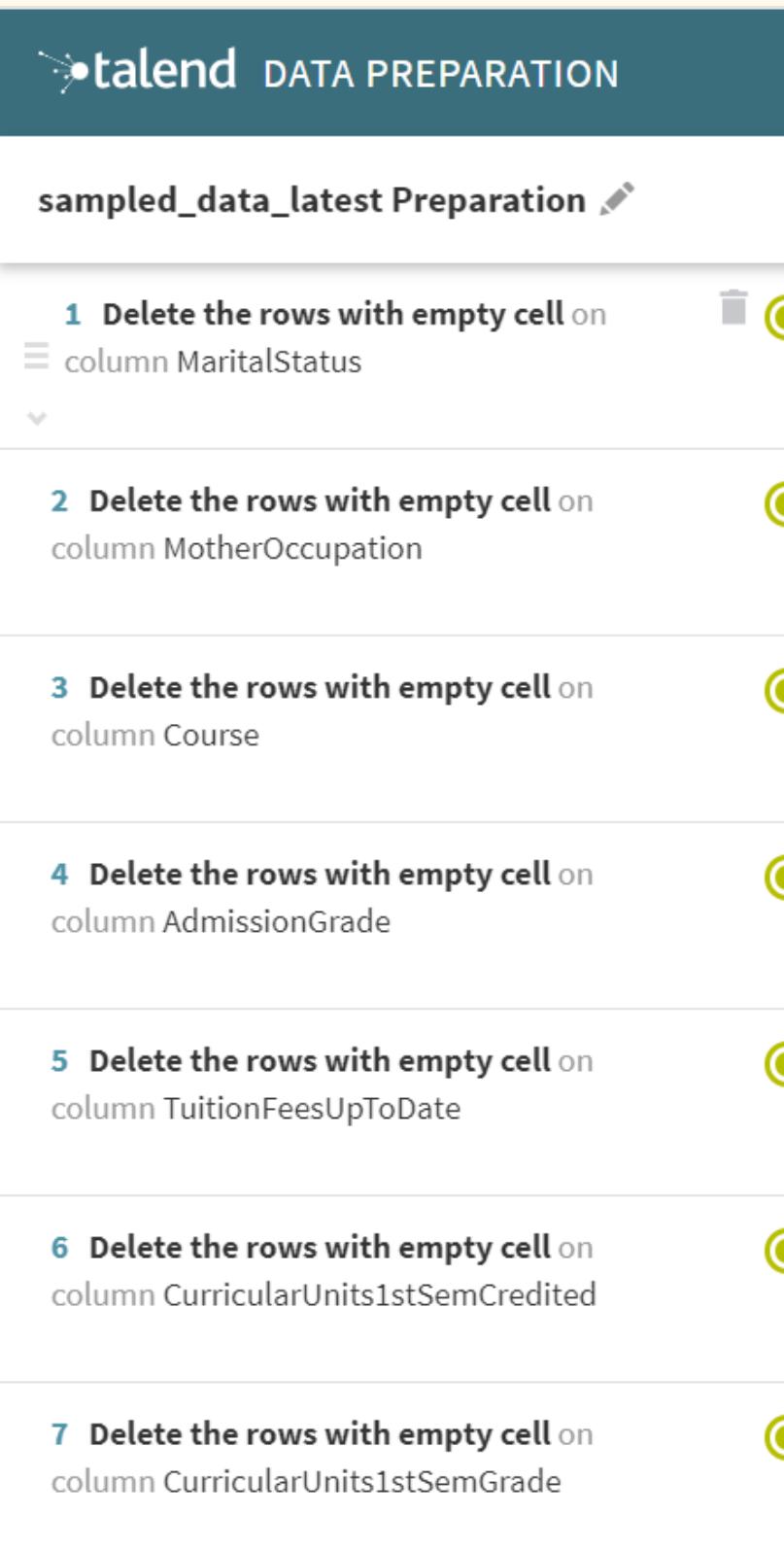
## Academic Outcomes By Courses



# MODIFY

# Data Preprocessing

- Handle missing values



# Data Preprocessing

- Encode categorical variables

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**11 Replace the cells that match on column Target**

**12 Replace the cells that match on column Target**

**13 Replace the cells that match on column Target**

Current:  $\cong$  Dropout

Replacement: 1

Overwrite entire cell

**SUBMIT**

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**11 Replace the cells that match on column Target**

**12 Replace the cells that match on column Target**

**13 Replace the cells that match on column Target**

Current:  $\cong$  Dropout

Replacement: 1

Overwrite entire cell

**SUBMIT**

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**11 Replace the cells that match on column Target**

**12 Replace the cells that match on column Target**

**13 Replace the cells that match on column Target**

Current:  $\cong$  Enrolled

Replacement: 2

Overwrite entire cell

**SUBMIT**

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**11 Replace the cells that match on column Target**

**12 Replace the cells that match on column Target**

**13 Replace the cells that match on column Target**

Current:  $\cong$  Graduate

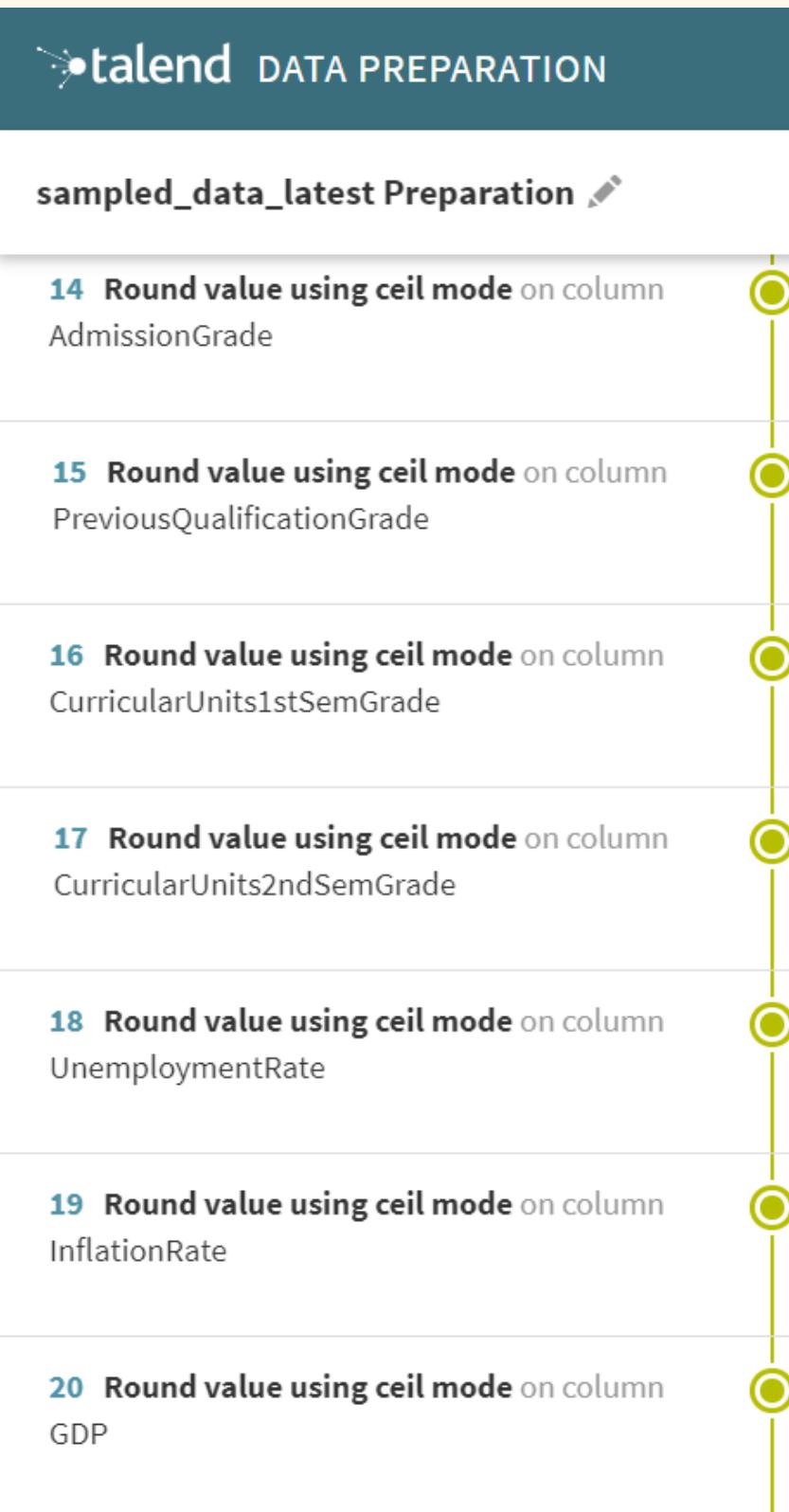
Replacement: 3

Overwrite entire cell

**SUBMIT**

# Data Preprocessing

- Standardization of data



# Data Preprocessing

- Perform feature engineering

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**22 Add, multiply, subtract or divide on**  
column CurricularUnits1stSemGrade\_copy

Operator: **+**

Use with: **Other column**

Column: **CurricularUnits2ndSemGrade**

**SUBMIT**

**talend DATA PREPARATION**

**sampled\_data\_latest Preparation**

**24 Add, multiply, subtract or divide on**  
column CurricularUnits1stSemGrade\_copy ...

Operator: **/**

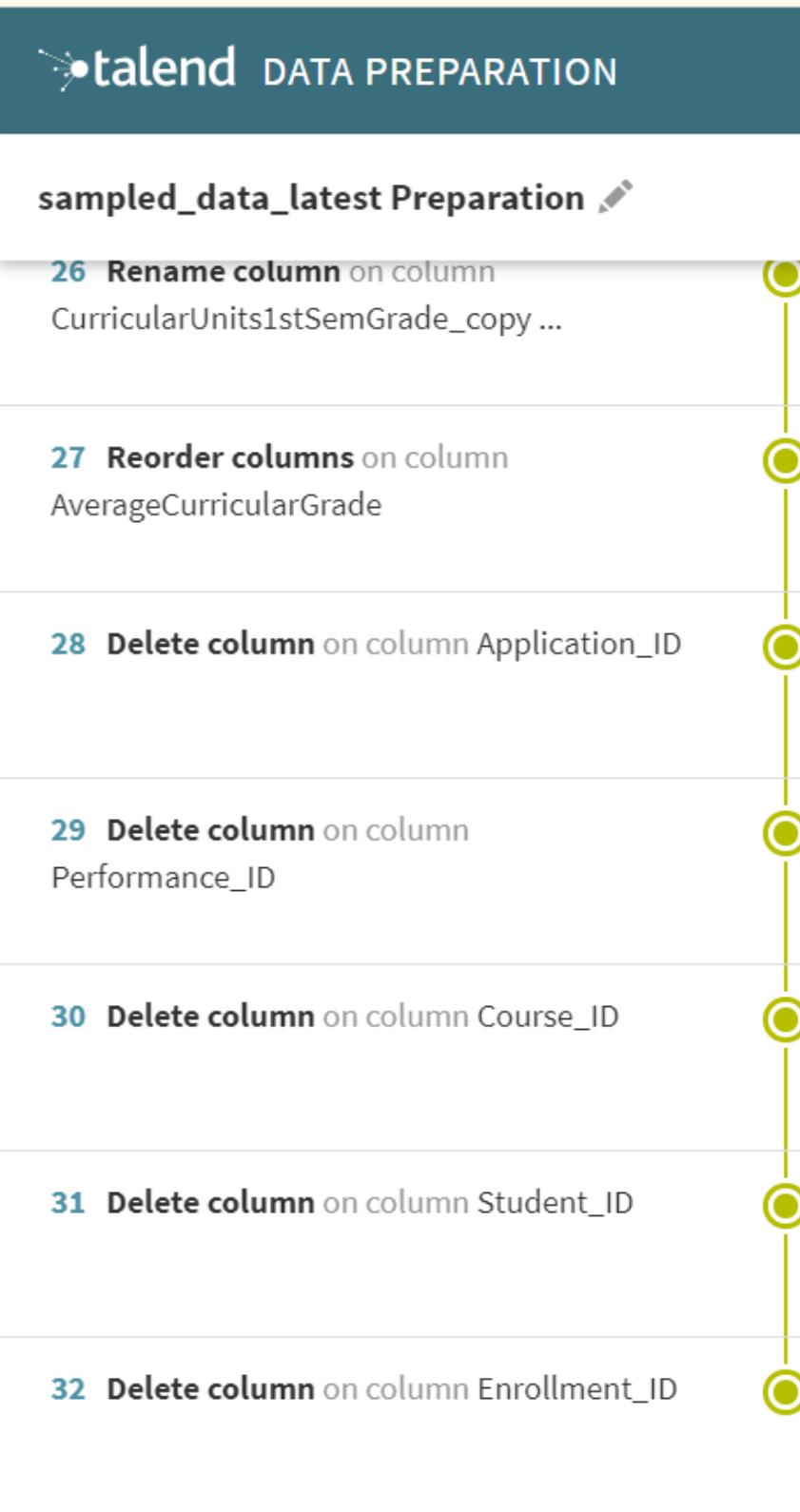
Use with: **Value**

Operand: **2**

**SUBMIT**

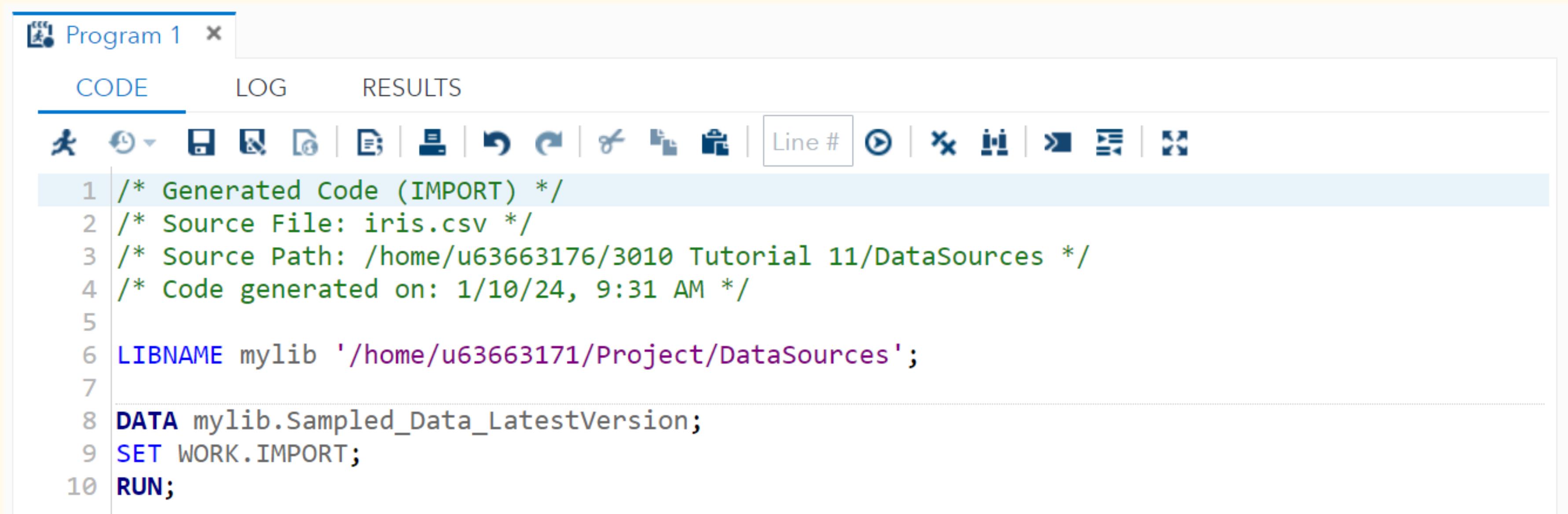
# Data Preprocessing

- Remove unused identifiers



# Data Preprocessing

- Code to convert the uploaded dataset to sas7bdat format



The screenshot shows a SAS program editor window titled "Program 1". The "CODE" tab is selected, displaying the following SAS code:

```
1 /* Generated Code (IMPORT) */
2 /* Source File: iris.csv */
3 /* Source Path: /home/u63663176/3010 Tutorial 11/DataSources */
4 /* Code generated on: 1/10/24, 9:31 AM */

5
6 LIBNAME mylib '/home/u63663171/Project/DataSources';
7
8 DATA mylib.Sampled_Data_LatestVersion;
9 SET WORK.IMPORT;
10 RUN;
```

The code is used to import the "iris.csv" dataset into a SAS library named "mylib". The imported dataset is named "Sampled\_Data\_LatestVersion". The code was generated on January 10, 2024, at 9:31 AM.

# Data Preprocessing

- Splitting of data using Data Partition node



# Data Preprocessing

- Data Partition Summary

## Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS2.Ids_DATA	1317
TRAIN	EMWS2.Part_TRAIN	922
VALIDATE	EMWS2.Part_VALIDATE	395

# **Techniques And Algorithm**

# Association Rule

## Rule Description

Map	Rule
RULE1	$2 \Rightarrow 1$
RULE2	$3 \Rightarrow 1$
RULE3	$3 \& 2 \Rightarrow 1$
RULE4	$1 \Rightarrow 2$
RULE5	$3 \Rightarrow 2$
RULE6	$3 \& 1 \Rightarrow 2$
RULE7	$3 \Rightarrow 2 \& 1$
RULE8	$2 \Rightarrow 3$
RULE9	$1 \Rightarrow 3$
RULE10	$2 \& 1 \Rightarrow 3$
RULE11	$2 \Rightarrow 3 \& 1$
RULE12	$1 \Rightarrow 3 \& 2$

# Association Rule

## Rule Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count
2	100.00	100.00	100.00	1.00	17.00
2	100.00	100.00	94.12	1.00	16.00
3	100.00	100.00	94.12	1.00	16.00
2	100.00	100.00	100.00	1.00	17.00
2	100.00	100.00	94.12	1.00	16.00
3	100.00	100.00	94.12	1.00	16.00
3	100.00	100.00	94.12	1.00	16.00
2	94.12	94.12	94.12	1.00	16.00
2	94.12	94.12	94.12	1.00	16.00
3	94.12	94.12	94.12	1.00	16.00
3	94.12	94.12	94.12	1.00	16.00
3	94.12	94.12	94.12	1.00	16.00

Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Index	Transpose Rule
2 ==> 1	2	1	2	=====>	1		1	1
3 ==> 1	3	1	3	=====>	1		2	1
3 & 2 ==> 1	3 & 2	1	3	2	=====>	1	3	1
1 ==> 2	1	2	1	=====>	2		4	1
3 ==> 2	3	2	3	=====>	2		5	1
3 & 1 ==> 2	3 & 1	2	3	1	=====>	2	6	1
3 ==> 2 & 1	3	2 & 1	3	=====>	2	1	7	1
2 ==> 3	2	3	2	=====>	3		8	1
1 ==> 3	1	3	1	=====>	3		9	1
2 & 1 ==> 3	2 & 1	3	2	1	=====>	3	10	1
2 ==> 3 & 1	2	3 & 1	2	=====>	3	1	11	1
1 ==> 3 & 2	1	3 & 2	1	=====>	3	2	12	1

# Sequence Analysis

## Configuration

- Dataset Role: Transaction
- ID: Course
- Target: Target (Nominal)



## Results

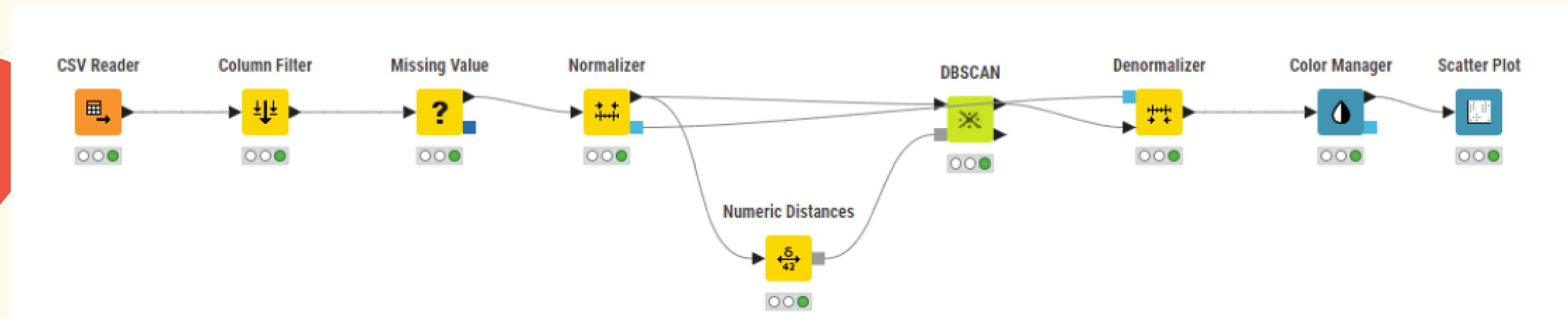
- Avg Expected Confidence & Confidence => 97.549%
- Avg support metric => 95.08%
- Lift Values => 1.

Rule Statistics					
The MEANS Procedure					
Variable	Label	Minimum	Maximum	Mean	
EXP_CONF	Expected Confidence(%)	94.1176471	100.0000000	97.5490196	
CONF	Confidence(%)	94.1176471	100.0000000	97.5490196	
SUPPORT	Support(%)	94.1176471	100.0000000	95.0980392	
LIFT	Lift	1.0000000	1.0000000	1.0000000	

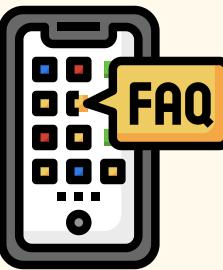


# DBSCAN

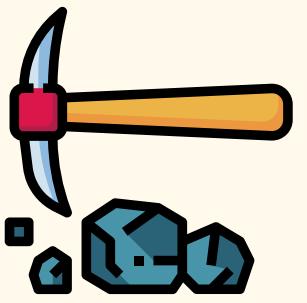
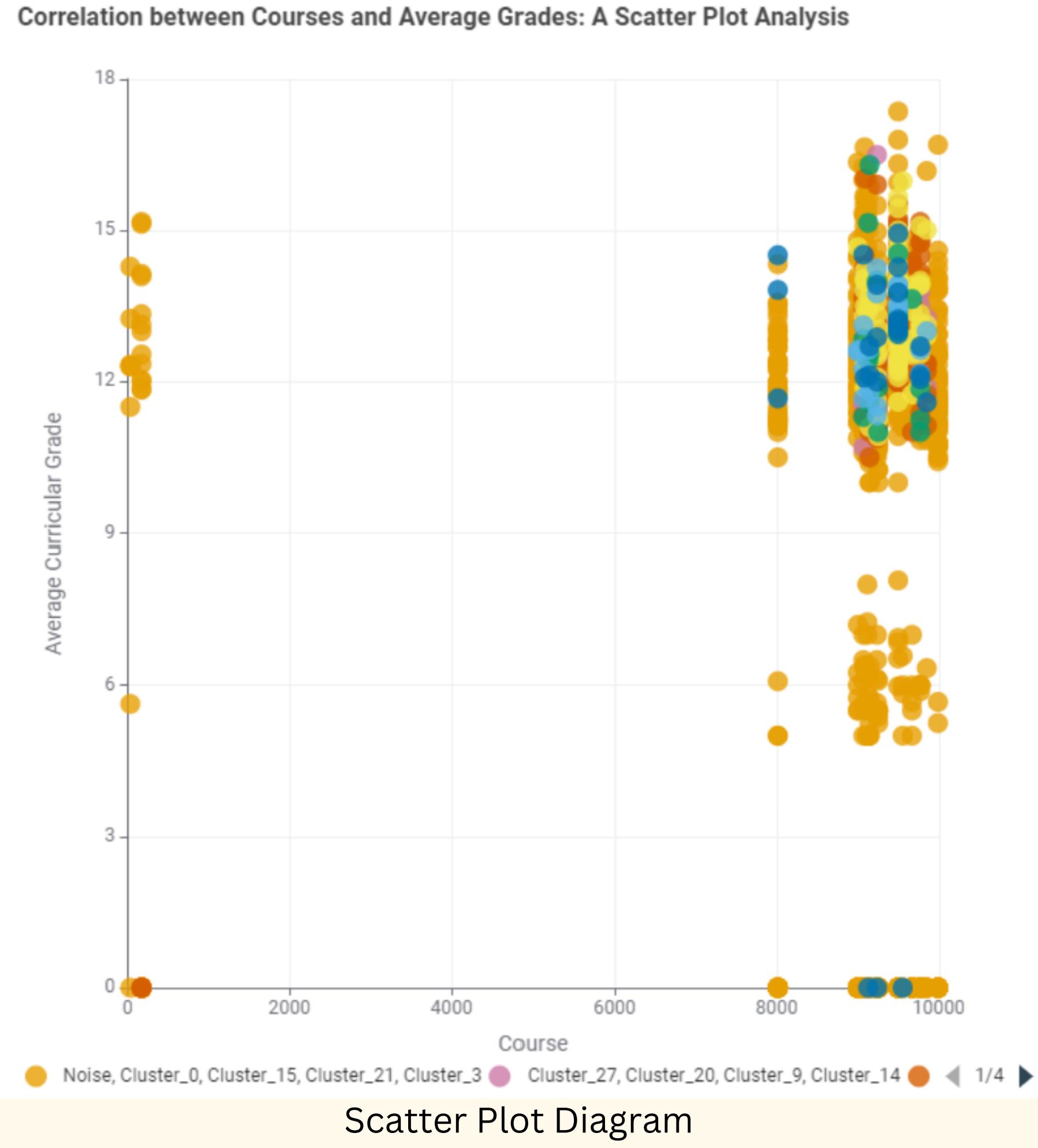
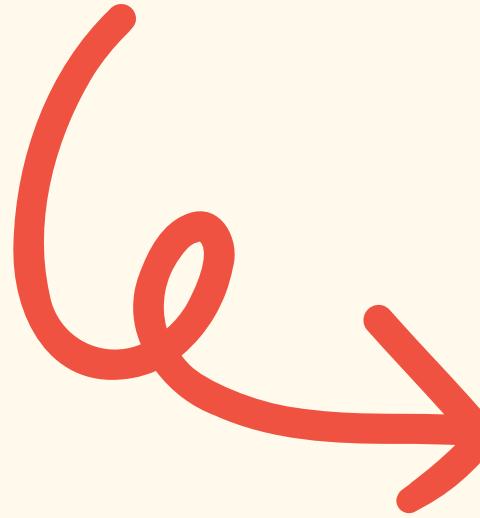
(Density-Based Spatial Clustering of Applications with Noise)



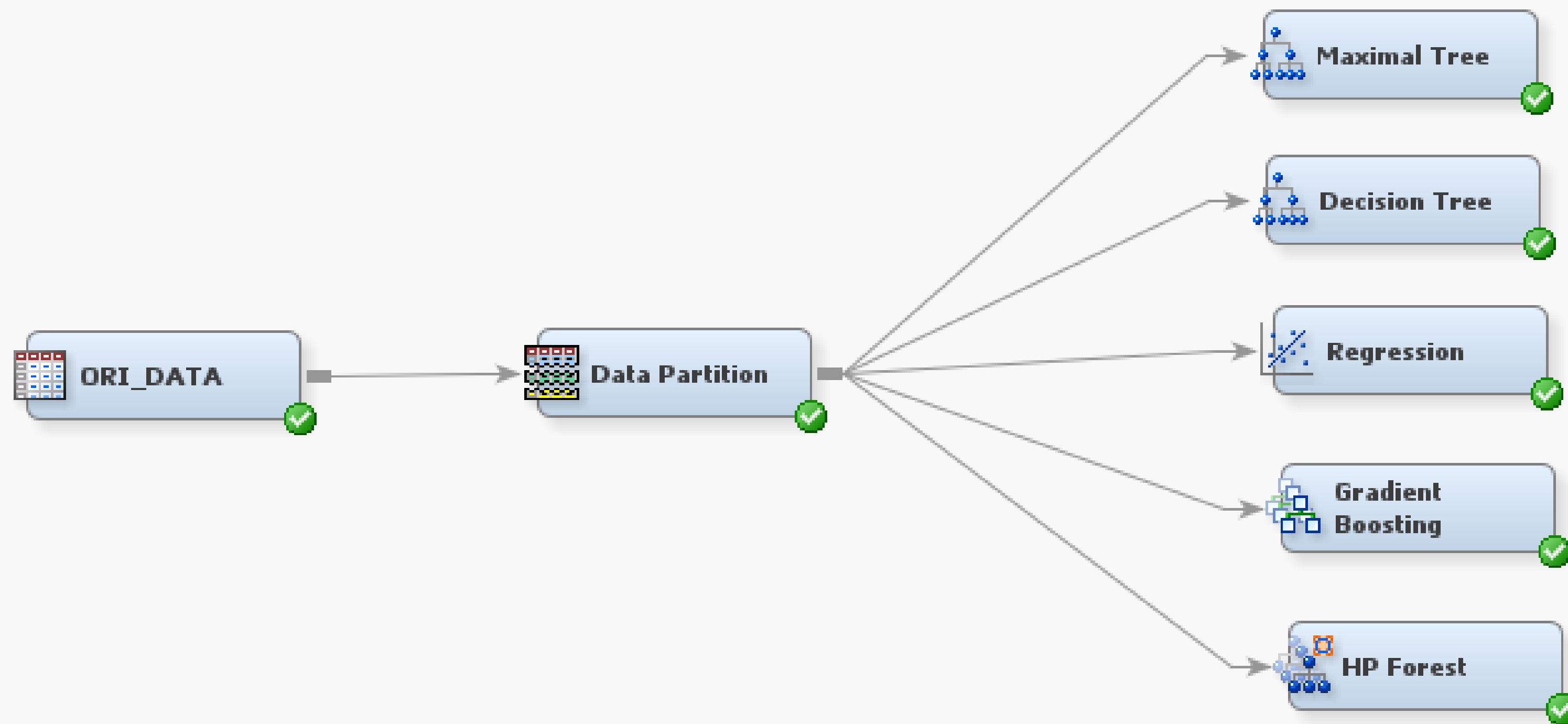
DBSCAN Diagram



# DBSCAN



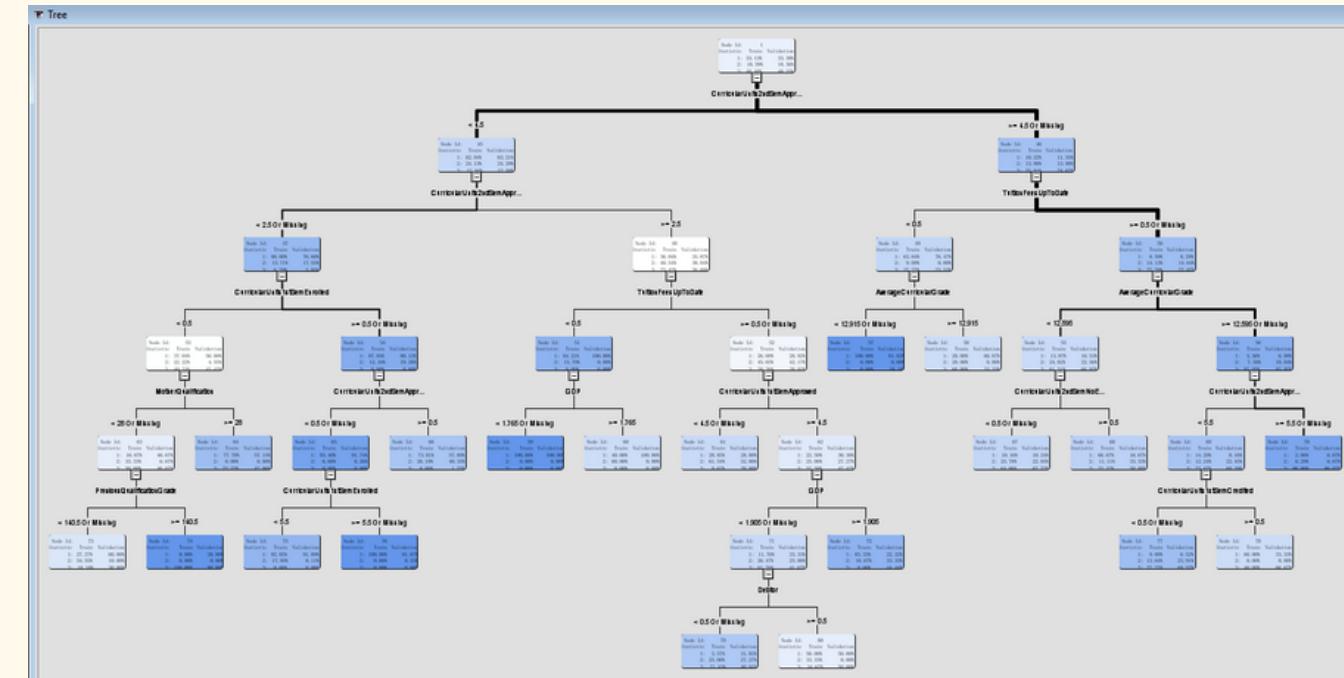
# Model



# Decision Tree (Maximal Tree)

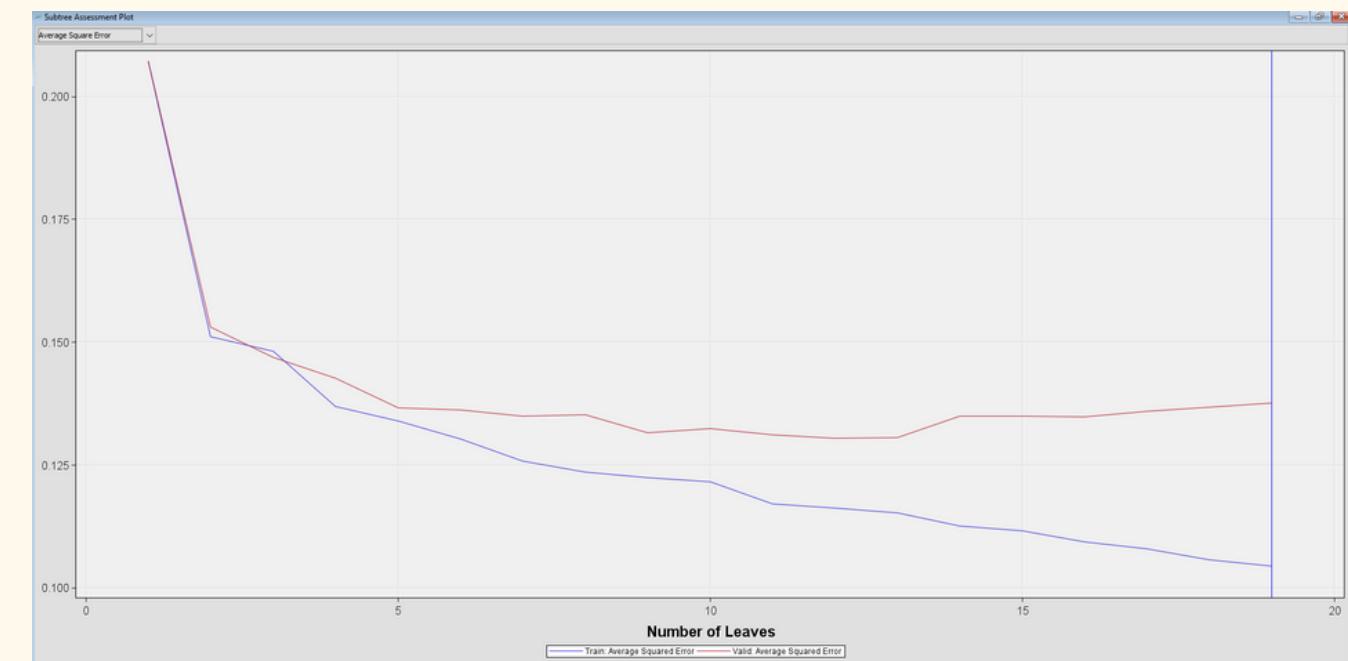
# Tree

- Selecting highest log worth input to be split.
  - CurricularUnits2ndSemApproved, log worth = 59.0071.
  - Left branch:  $< 4.5$ , Target = 1.
  - Right branch:  $\geq 4.5$ , Target = 3.
  - Repeat the splitting and train the root node.



# Misclassification Rate

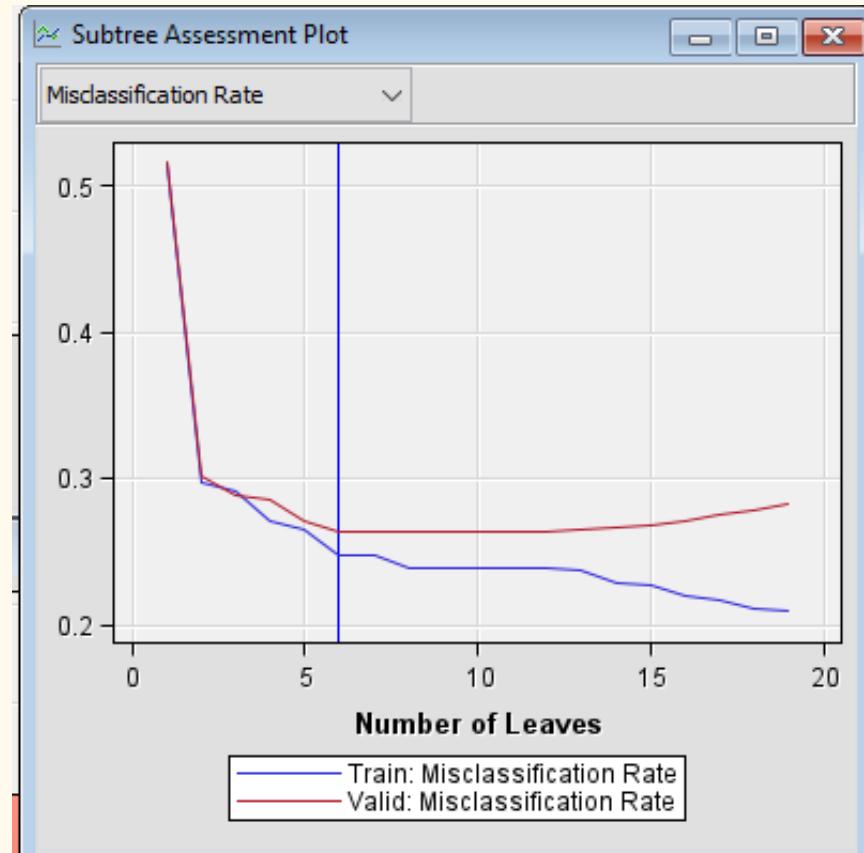
- Performance of validation sample improves up to a tree with approximately 5 leaves.
  - Become worse as complexity increases.
  - Overfitting - accurate predictions for test data but not valid data.



# Decision Tree (Optimal Tree)

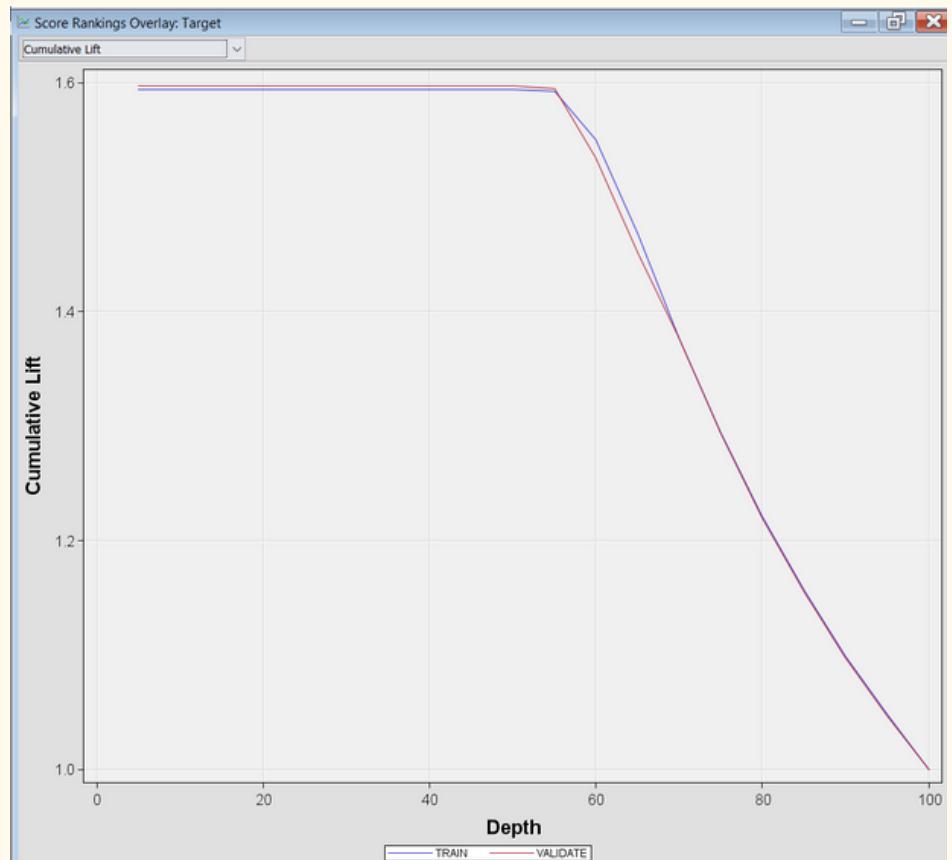
## Misclassification Rate

- 6-leaf tree has lowest average squared error.



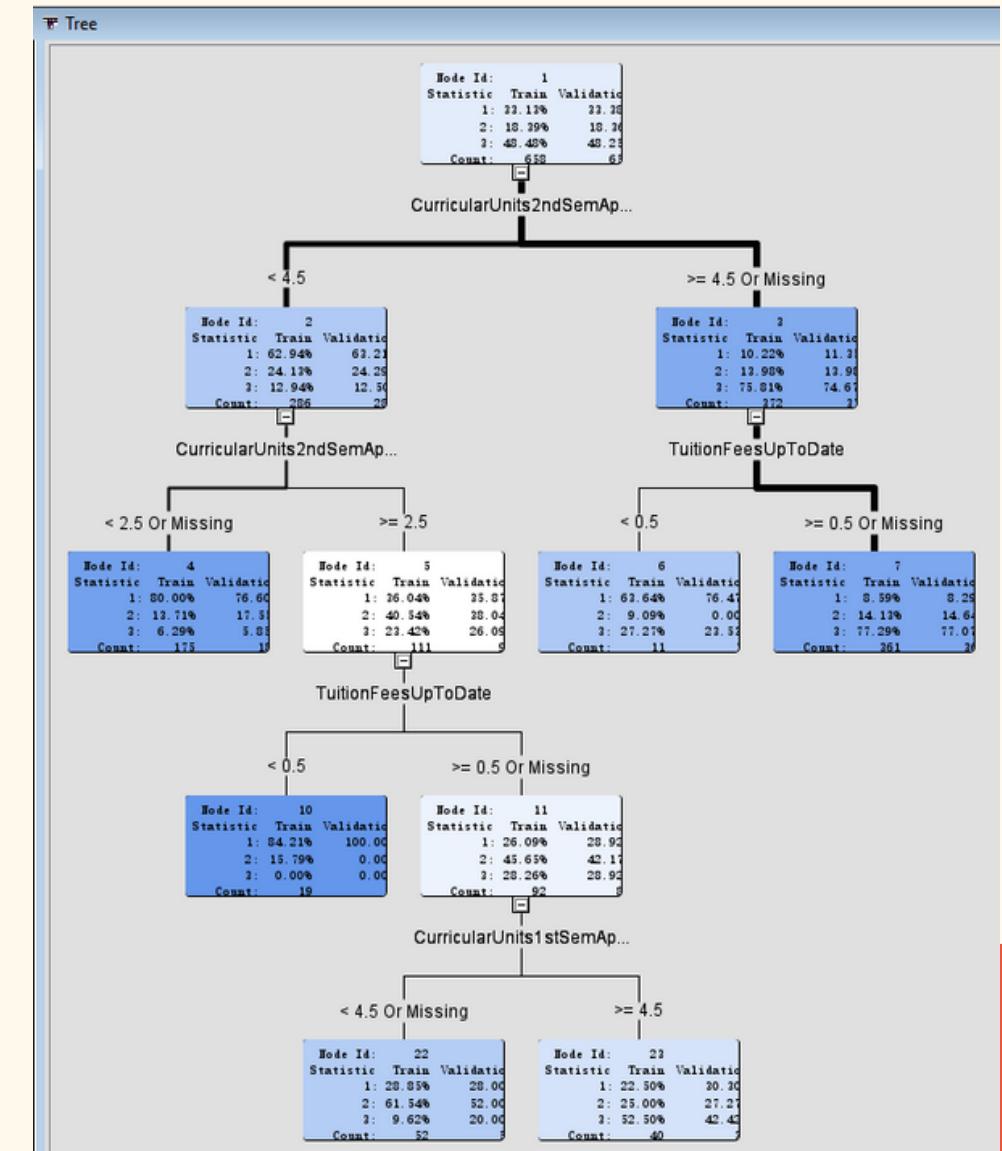
## Cumulative Lift

- First 50%, training and validation have high value of 1.6.
- Decline gradually as moving down the depth axis.
- Model become less effective when the decision tree depth goes deeper.



## Tree

Method: Assessment  
Measure: Decision



# Regression

## Summary

- All variables are statistically significant except MotherOccupation
- Order of variable suggests their relative importance in predicting outcome variables.

## Odds Ratio

- AgeAtEnrollment increase, chances for Target 2 and 3 to get 'Yes' decreases because the odds ration < 1.

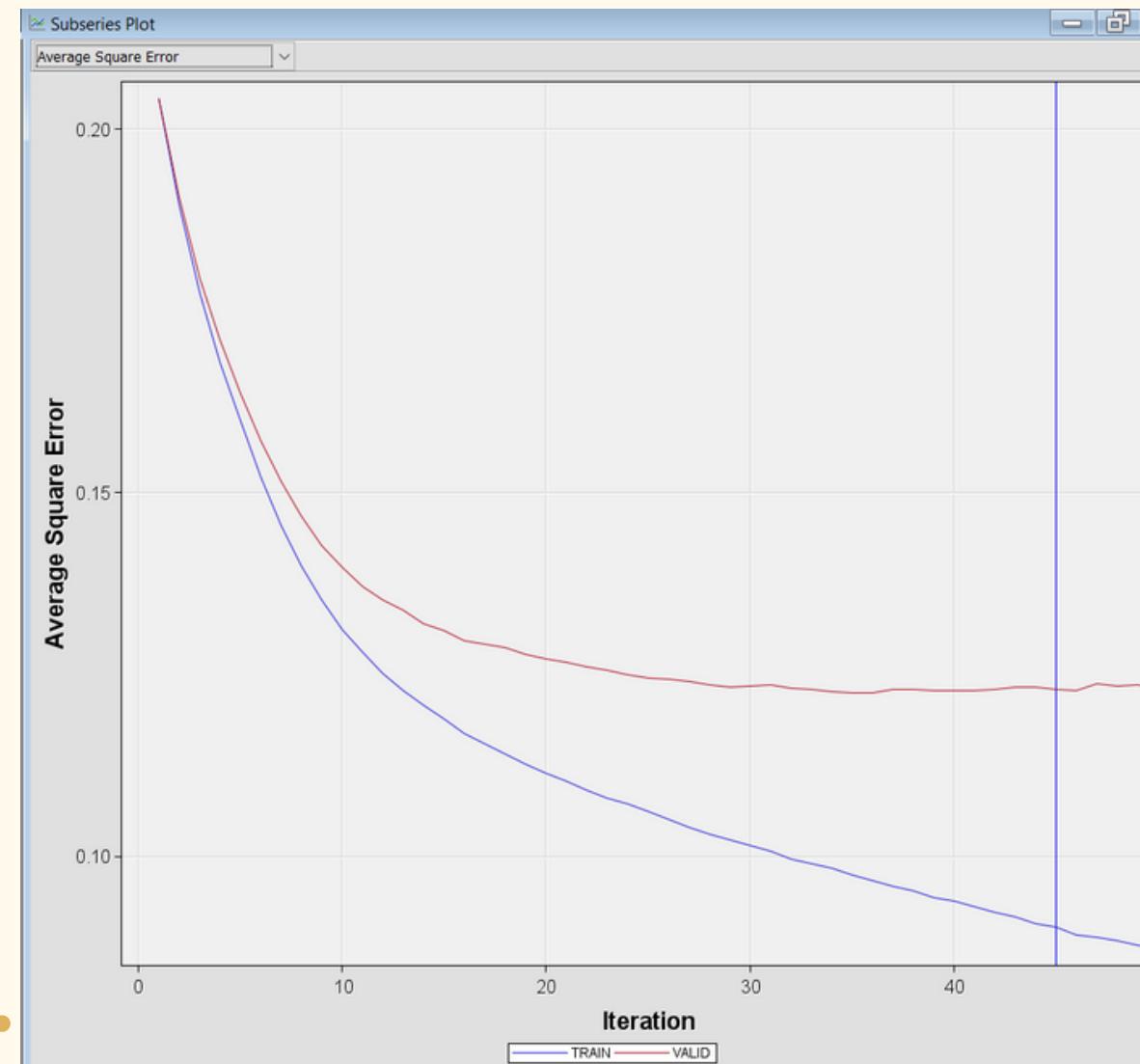
Summary of Stepwise Selection								
Step	Effect		DF	Number	Score	Wald	Validation	
	Entered	Removed					In	Chi-Square
1	CurricularUnits2ndSemApproved		2	1	262.8973		<.0001	1053.1
2	CurricularUnits2ndSemEnrolled		2	2	148.8320		<.0001	946.4
3	TuitionFeesUpToDate		2	3	47.2663		<.0001	876.8
4	AgeAtEnrollment		2	4	24.3491		<.0001	890.8
5	CurricularUnits1stSemApproved		2	5	19.1507		<.0001	874.6
6	CurricularUnits1stSemCredited		2	6	15.6976		0.0004	861.5
7	ScholarshipHolder		2	7	10.2562		0.0059	857.1
8	Gender		2	8	8.6783		0.0130	855.2
9	CurricularUnits2ndSemEvaluations		2	9	7.4784		0.0238	846.9
10	MotherOccupation		2	10	6.9736		0.0306	850.2
11	MotherOccupation	MotherOccupation	2	9	5.4080		0.0669	846.9

Odds Ratio Estimates		
Effect	Target	Point Estimate
AgeAtEnrollment	3	0.939
AgeAtEnrollment	2	0.919
CurricularUnits1stSemApproved	3	1.901
CurricularUnits1stSemApproved	2	1.023
CurricularUnits1stSemCredited	3	0.761
CurricularUnits1stSemCredited	2	0.882
CurricularUnits2ndSemApproved	3	3.308
CurricularUnits2ndSemApproved	2	1.642
CurricularUnits2ndSemEnrolled	3	0.276
CurricularUnits2ndSemEnrolled	2	0.631
CurricularUnits2ndSemEvaluations	3	0.994
CurricularUnits2ndSemEvaluations	2	1.115
Gender	3	0.390
Gender	2	0.669
ScholarshipHolder	3	2.605
ScholarshipHolder	2	1.080
TuitionFeesUpToDate	3	46.915
TuitionFeesUpToDate	2	10.215

# Gradient Boosting

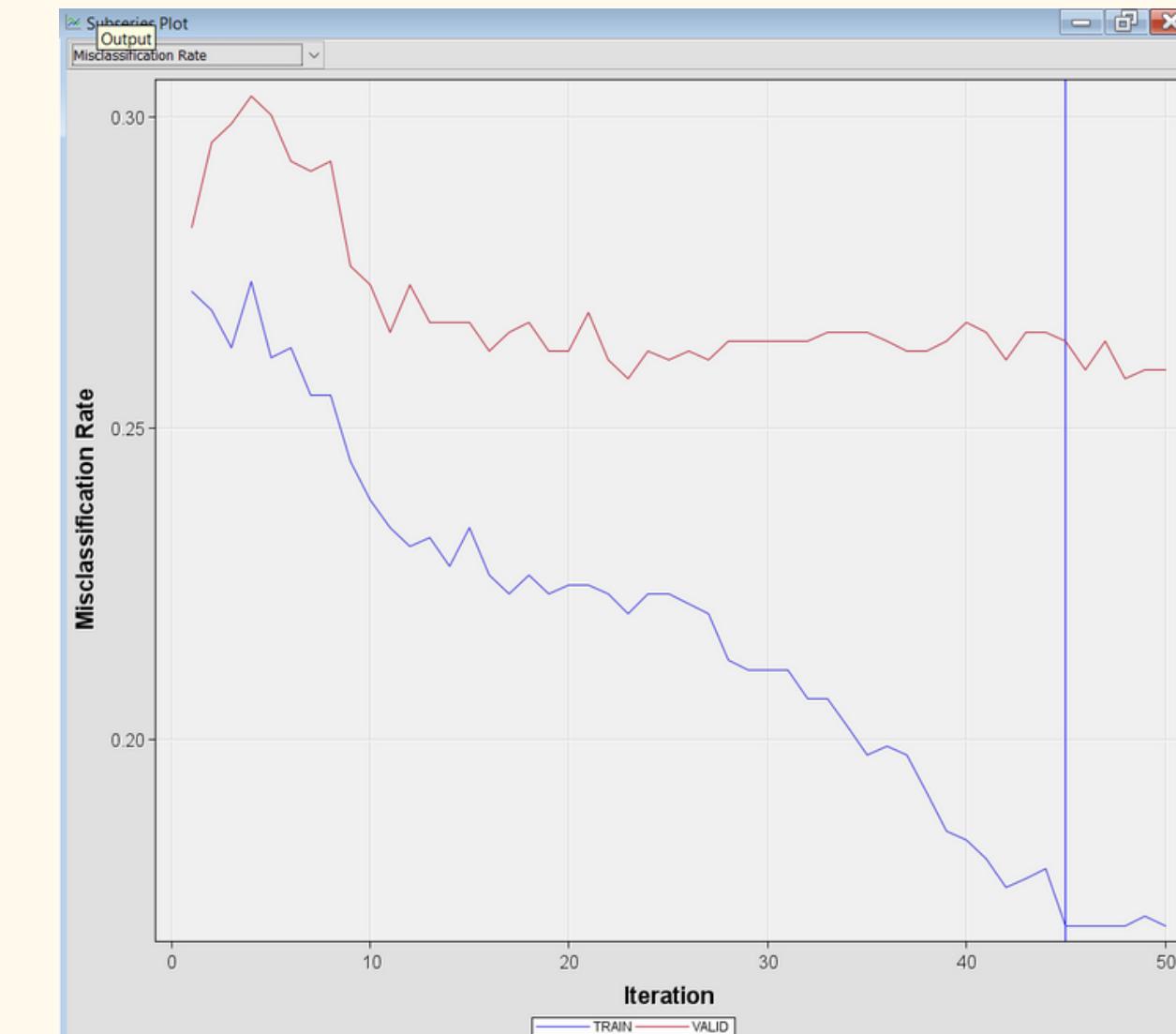
## Average Square Error

- Decreases significantly when depth < 10
- Should use tree with smaller depth in order to predict accurately.



## Misclassification Rule

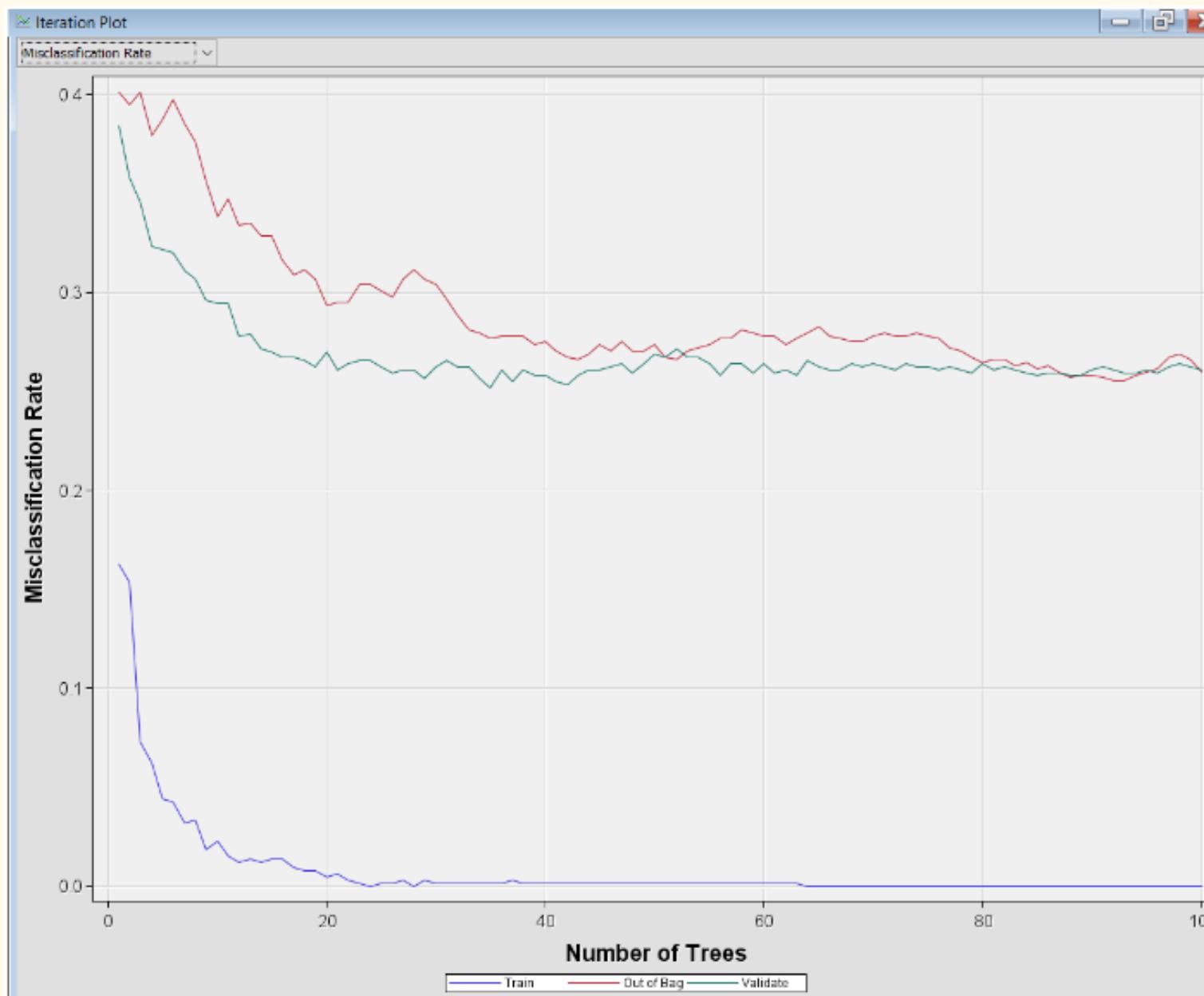
- When the training error decreases, the validation error increases.
- Find the optimal tree depth and stop the iteration before overfitting occurs



# Random Forest

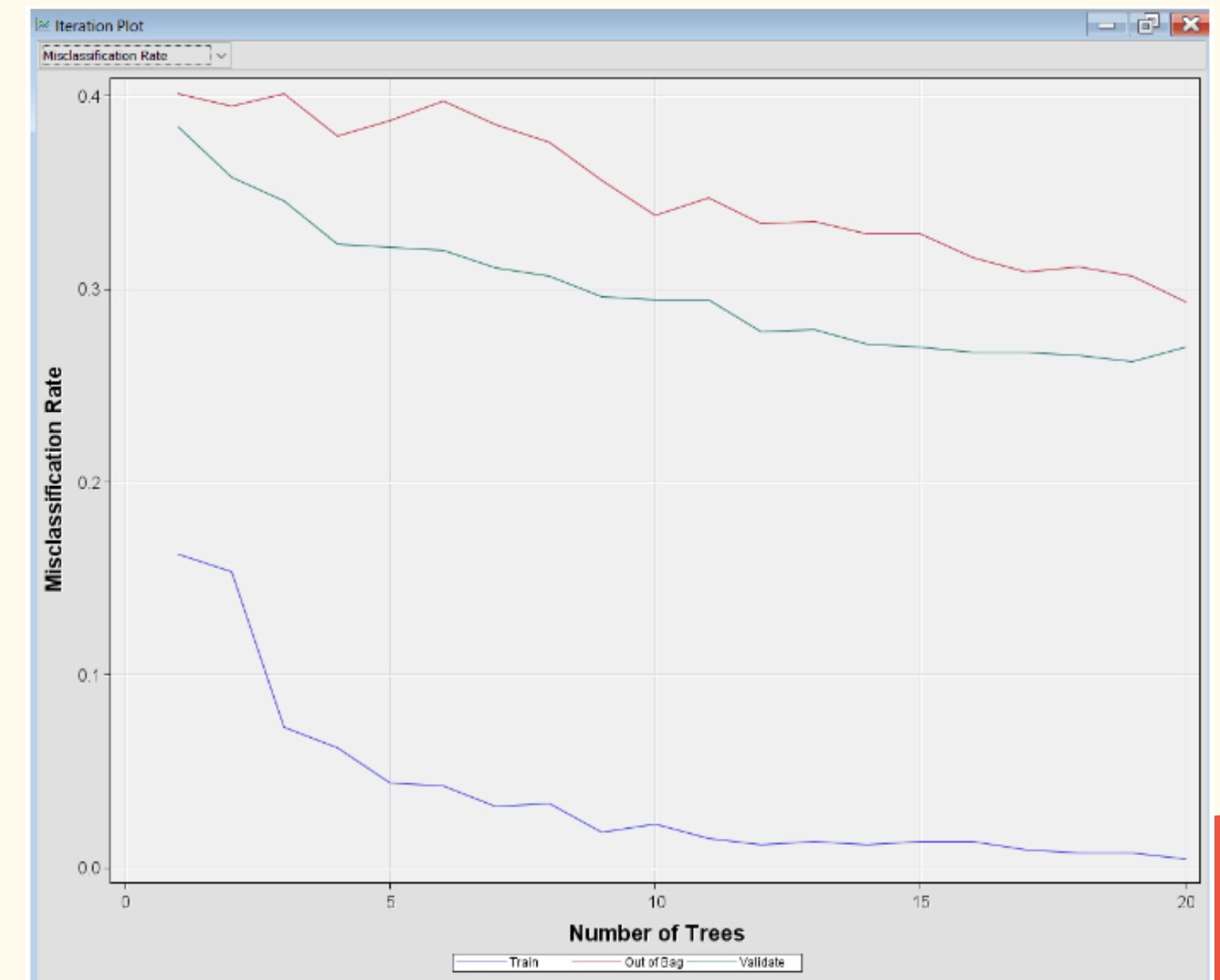
## Before Reduction

- Out-of-Bags (OOB) decreases as the number of trees increases.
- Train data misclassification plot started to flatten after 20 trees.

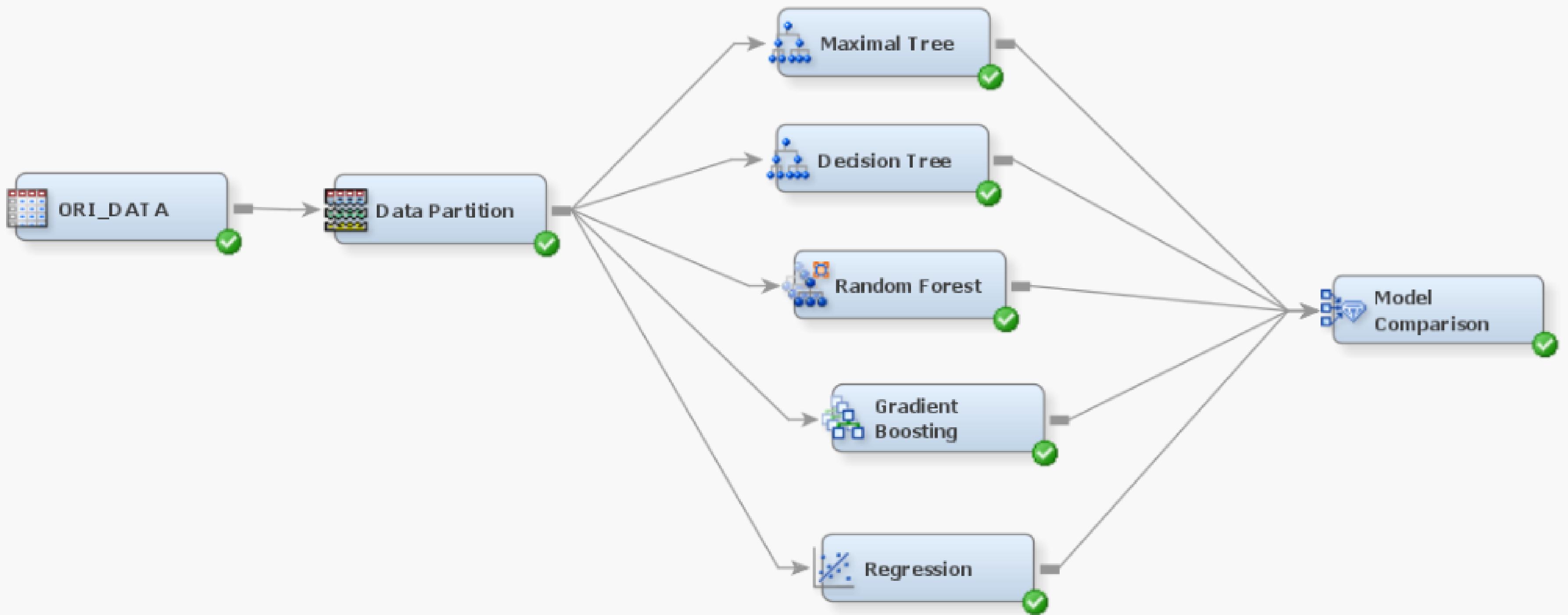


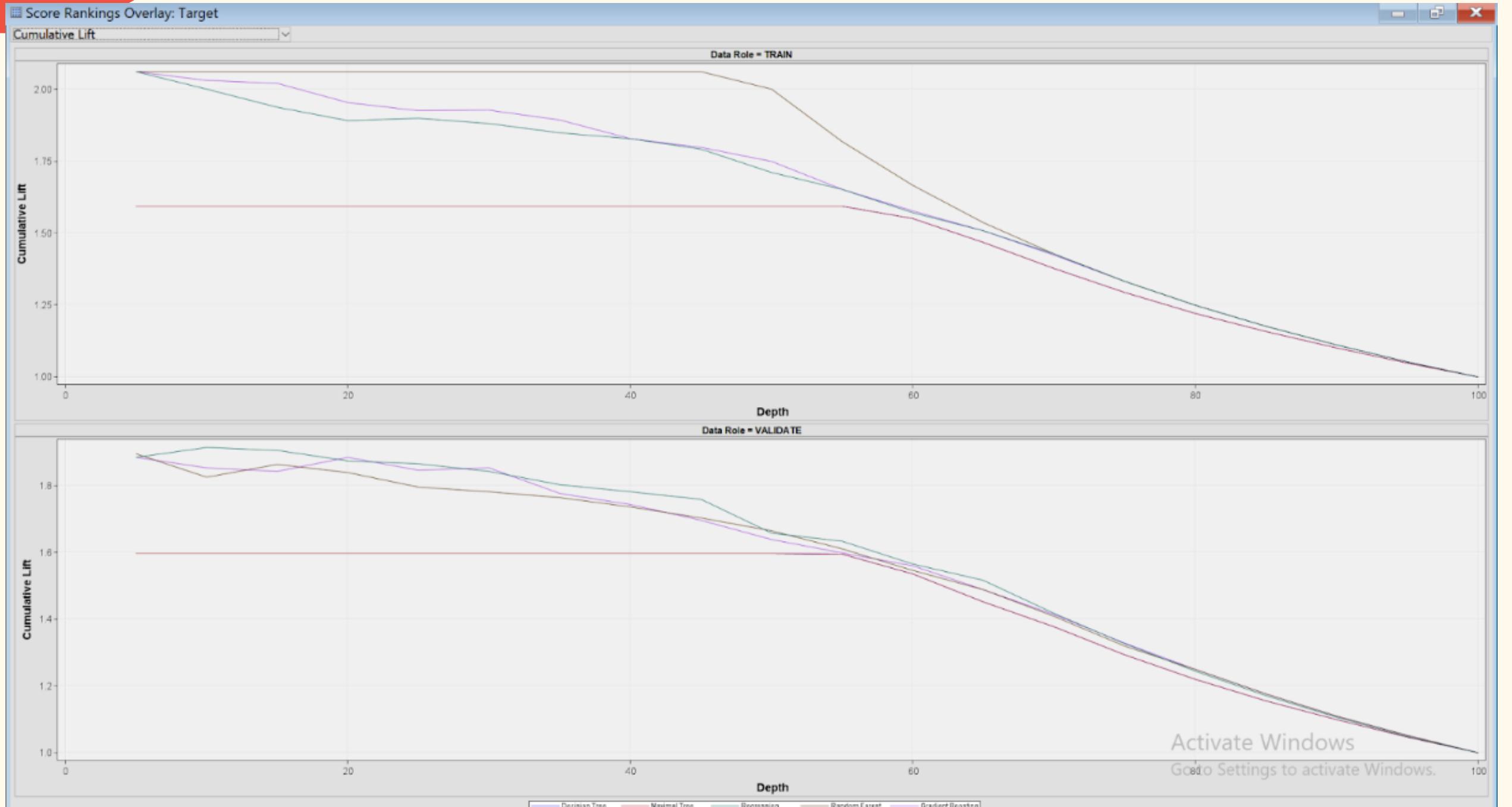
## After Reduction

- Misclassification Rate of validation data over time decreases.
- Not undergoing overfitting.



# Assess





## Cumulative Lift

- Random Forest model performs best on the training dataset.
- Regression model performs best on the validation dataset.
- All models perform similarly at greater depths.

# Fit Statistics Summary

- Regression Model
    - Lowest misclassification rate => 0.249
    - Akaike's Information Criterion (AIC) => 770.94
    - Highest ROC index => 0.909

### Event Classification Table

Model Selection based on Valid: Misclassification Rate (\_VMISC\_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Regression	TRAIN	Target	Target	24	256	83	295
Reg	Regression	VALIDATE	Target	Target	29	262	79	289
Tree	Maximal Tree	TRAIN	Target	Target	19	238	101	300
Tree	Maximal Tree	VALIDATE	Target	Target	25	239	102	293
Tree2	Decision Tree	TRAIN	Target	Target	19	238	101	300
Tree2	Decision Tree	VALIDATE	Target	Target	25	239	102	293
HPDMForest	Random Forest	TRAIN	Target	Target	.	339	.	319
HPDMForest	Random Forest	VALIDATE	Target	Target	25	245	96	293
Boost	Gradient Boosting	TRAIN	Target	Target	16	249	90	303
Boost	Gradient Boosting	VALIDATE	Target	Target	30	255	86	288

## Event Classification Table

- Regression model shows the smallest changes from training to validation
  - False negatives (24 ---> 29)
  - True positives (295 ---> 289).

# Conclusion

The research emphasizes the need for customized curriculum and support to lower dropout rates and assist financially challenged students.

It also showcases the crucial role of predictive models in early intervention.

Overall, it highlights how data mining can significantly improve educational strategies and student outcomes in higher education.



# Thank You

