WIE3007 DATA MINING AND WAREHOUSING

PRESENTATION VIDEO:

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

01 To investigate the factors that contribute to the academic dropouts.

02 To optimise the curriculum performance based on the identified factors.

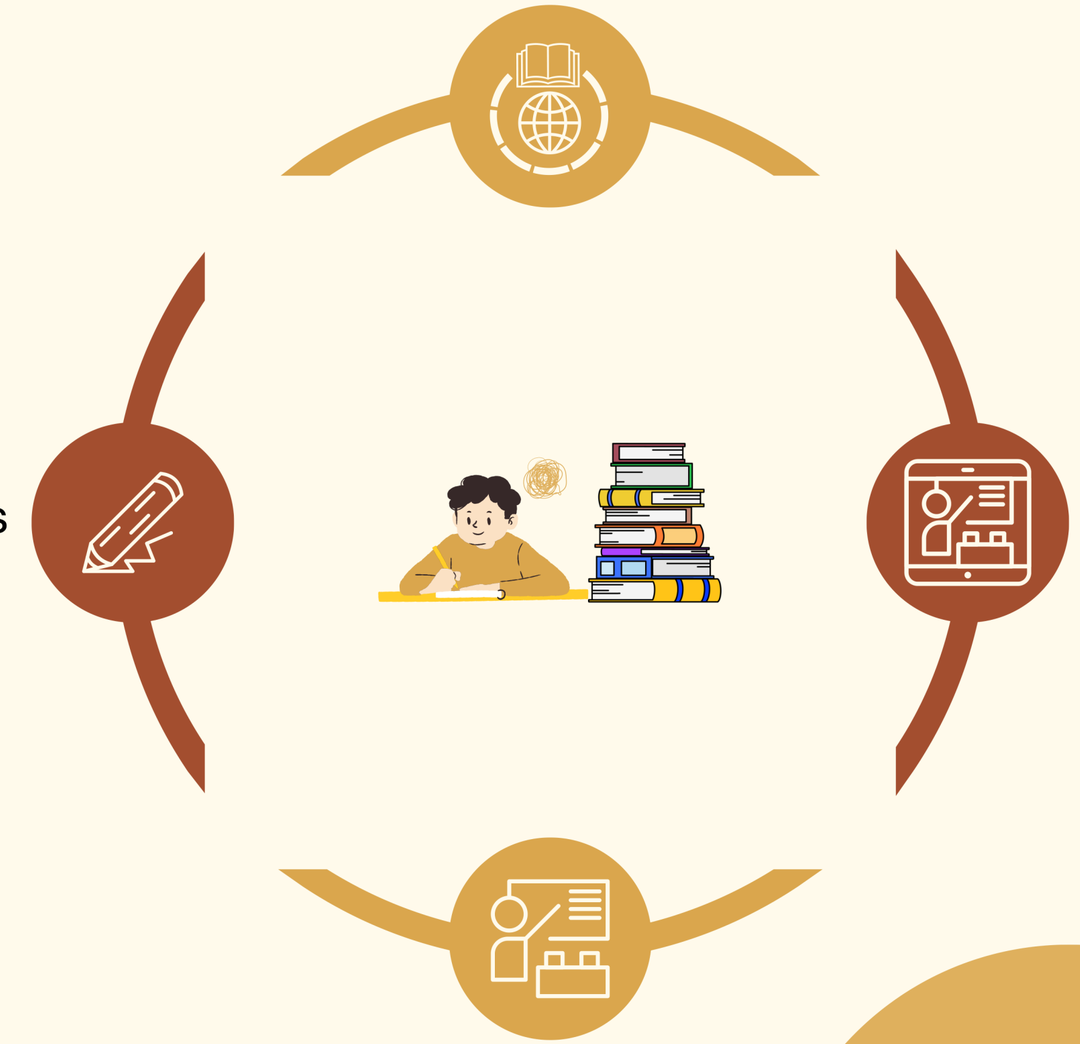
03 To foster a collaborative learning environment based on the factors that contribute to the academic success

Introduction

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

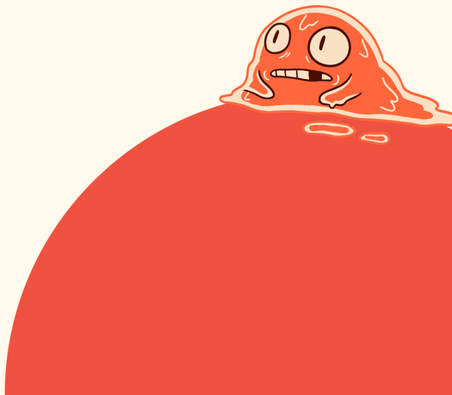
The dataset, comprising 4424 records and 36 features , is used for a three-category classification task (dropout, enrolment, and postgraduate).

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



Dataset source: Kaggle

Sample



Raw Data

	Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Fathe qualificati
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	
...	...	...	...	...	...	...	...	...	...	
4419	1	1	6	9773	1	1	125.0	1	1	
4420	1	1	2	9773	1	1	120.0	105	1	
4421	1	1	1	9500	1	1	154.0	1	37	
4422	1	1	1	9147	1	1	180.0	1	37	
4423	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\t	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Fathe qualificati
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		Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	...	Curricular units 2nd sem (credited)	Curri units (enro
1	3290	1	1	1	9238	1	1	125.0	1	1	1	...	0	
	3416	1	1	1	9773	1	1	143.0	1	3	3	...	0	
	2522	1	17	6	9500	1	1	131.0	1	1	19	...	0	
	4135	1	1	1	8014	0	1	115.0	1	38	38	...	0	
2	981	2	39	1	9991	0	1	100.0	1	19	19	...	0	
	2006	2	39	1	8014	0	1	146.0	1	37	37	...	0	
	4034	2	39	1	8014	0	19	133.1	1	34	34	...	0	
	1739	2	39	1	9070	1	19	133.1	1	37	37	...	0	
3	688	3	39	1	9991	0	1	120.0	1	37	1	...	0	
	166	3	39	1	9003	1	1	170.0	1	1	37	...	0	
	1428	3	17	1	9238	1	1	138.0	1	37	37	...	0	
	428	3	1	2	9085	1	1	135.0	1	19	1	...	0	
4	409	4	39	1	9085	1	1	133.1	1	34	34	...	0	
	939	4	39	1	8014	0	19	133.1	1	37	37	...	0	
	2999	4	39	1	8014	0	1	138.0	1	19	19	...	0	
	2794	4	39	1	9991	0	1	130.0	1	37	37	...	0	
5	4374	5	7	1	9500	1	40	150.0	1	37	37	...	1	
	3943	5	39	1	9147	1	1	180.0	1	1	1	...	0	
	3721	5	7	1	9070	1	3	140.0	1	37	37	...	0	
	2919	5	39	1	9003	1	19	133.1	1	19	1	...	0	
6	2914	6	39	2	9147	1	1	133.1	1	37	37	...	0	
	2915	6	39	1	8014	0	1	133.1	1	37	37	...	0	
	1180	6	39	1	9500	1	1	133.1	1	37	37	...	0	
	1572	6	39	1	9991	0	12	130.0	1	37	37	...	0	

24 rows × 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\t	Previous qualification	Previous qualification (grade)	Nacionality	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	

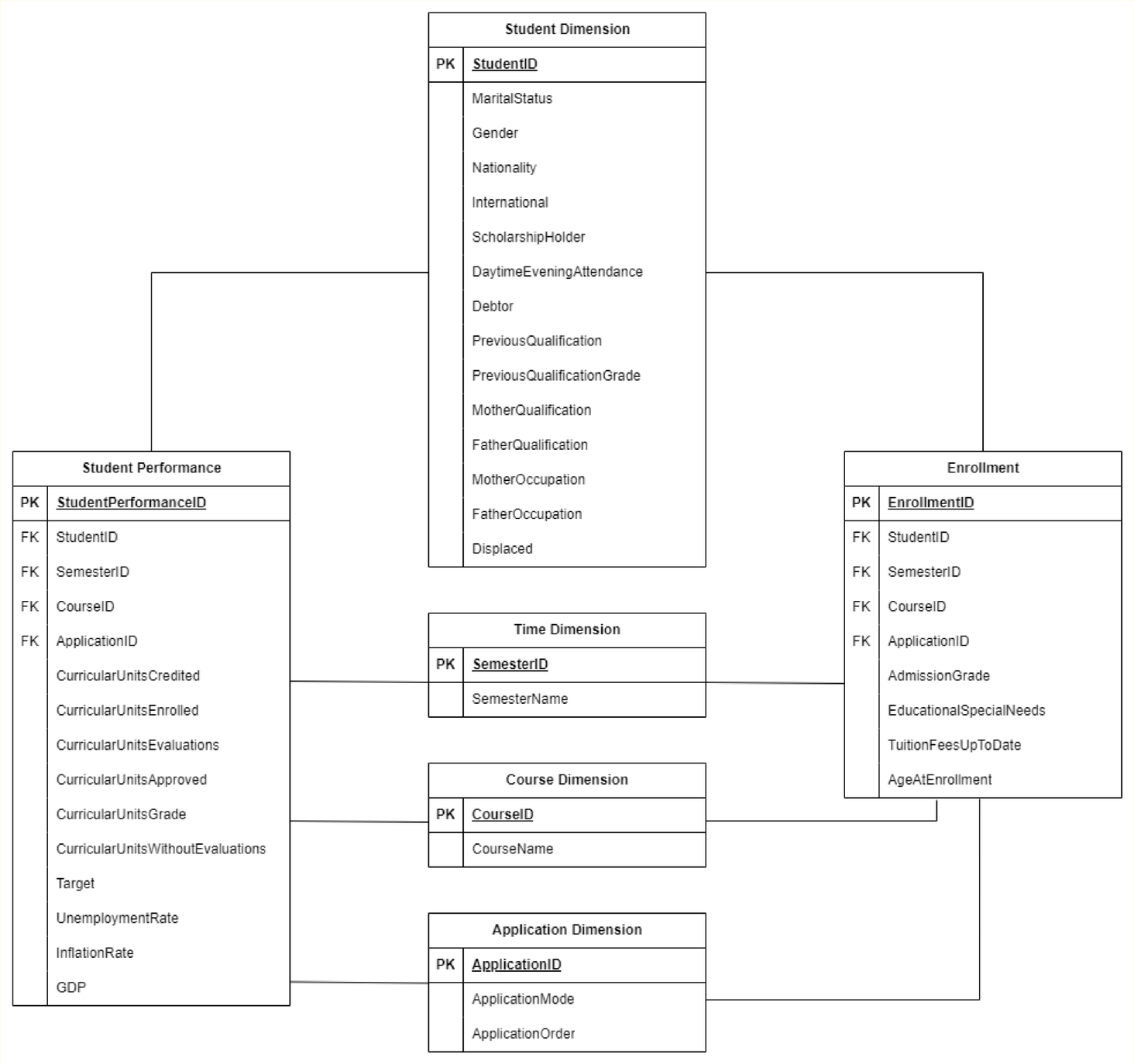
2212 rows × 37 columns

Featuretools And Star Schema

Featuretools

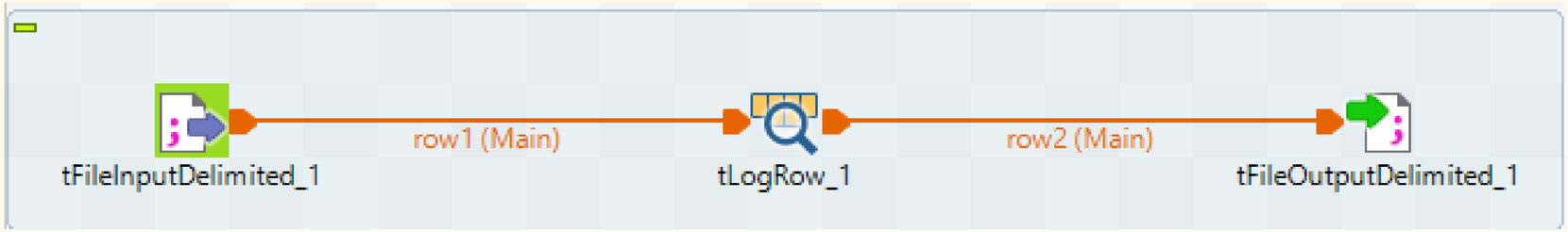
- Read dataset from csv file
- Define entities
- Create entities set
- Add dataframe to entity set
- Establish relationship
- Perform deep feature synthesis

Star Schema


Student Dimension
PK StudentID
MaritalStatus
Gender
Nationality
International
ScholarshipHolder
DaytimeEveningAttendance
Debtor
PreviousQualification
PreviousQualificationGrade
MotherQualification
FatherQualification
MotherOccupation
FatherOccupation
Displaced
Student Performance
PK StudentPerformanceID
FK StudentID
FK SemesterID
FK CourseID
FK ApplicationID
CurricularUnitsCredited
CurricularUnitsEnrolled
CurricularUnitsEvaluations
CurricularUnitsApproved
CurricularUnitsGrade
CurricularUnitsWithoutEvaluations
Target
UnemploymentRate
InflationRate
GDP
Time Dimension
PK SemesterID
SemesterName
Course Dimension
PK CourseID
CourseName
Application Dimension
PK ApplicationID
ApplicationMode
ApplicationOrder
Enrollment
PK EnrollmentID
FK StudentID
FK SemesterID
FK CourseID
FK ApplicationID
AdmissionGrade
EducationalSpecialNeeds
TuitionFeesUpToDate
AgeAtEnrollment

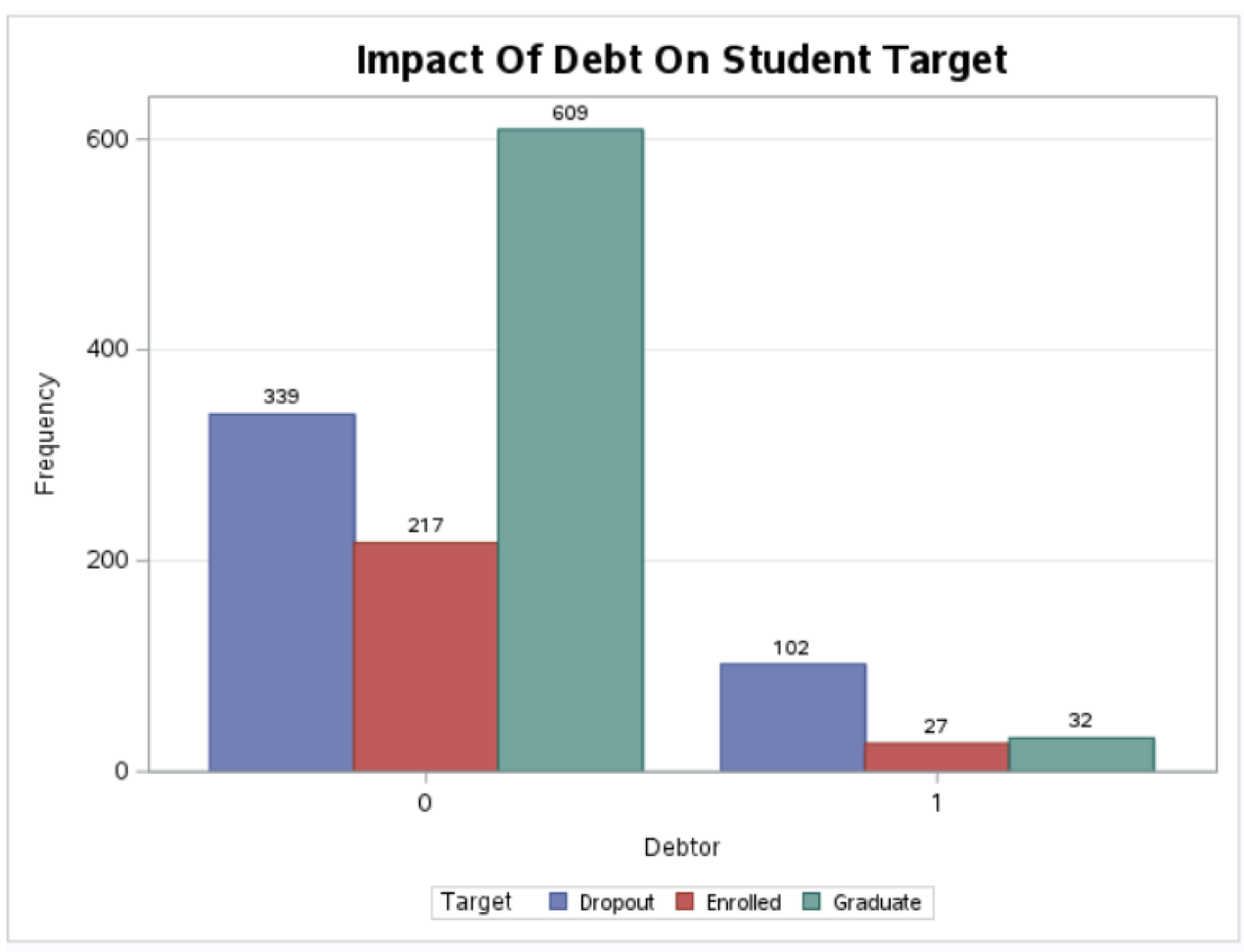
Explore

Extract Data From CSV File Using Talend Open Studio



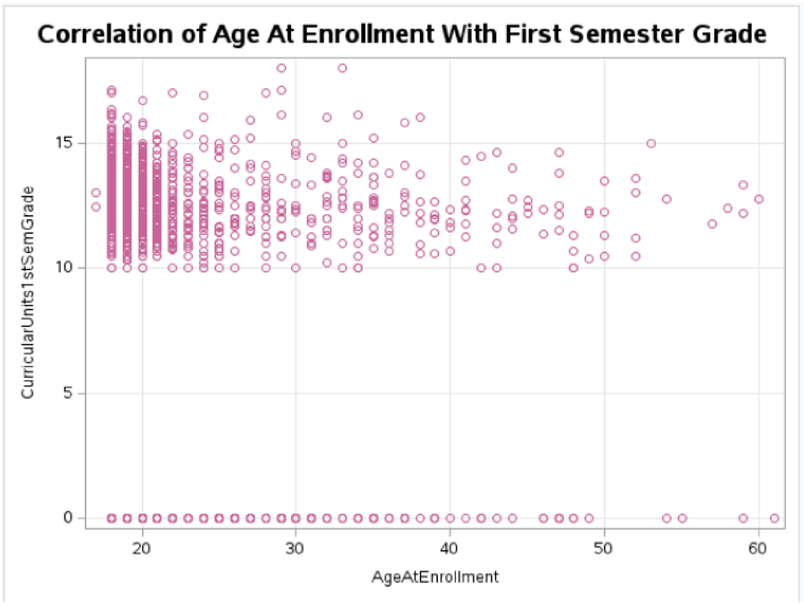
Exploratory Data Analysis

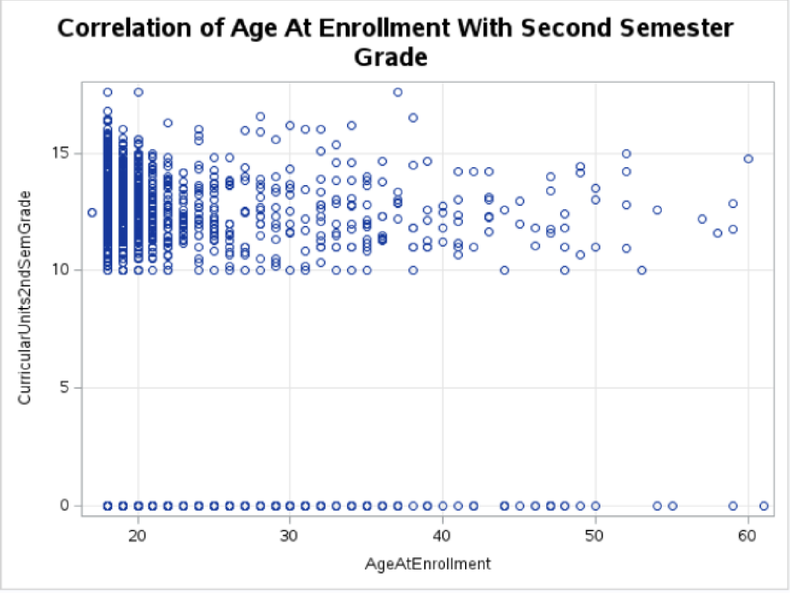
Impact Of Debt On Student Target



Exploratory Data Analysis

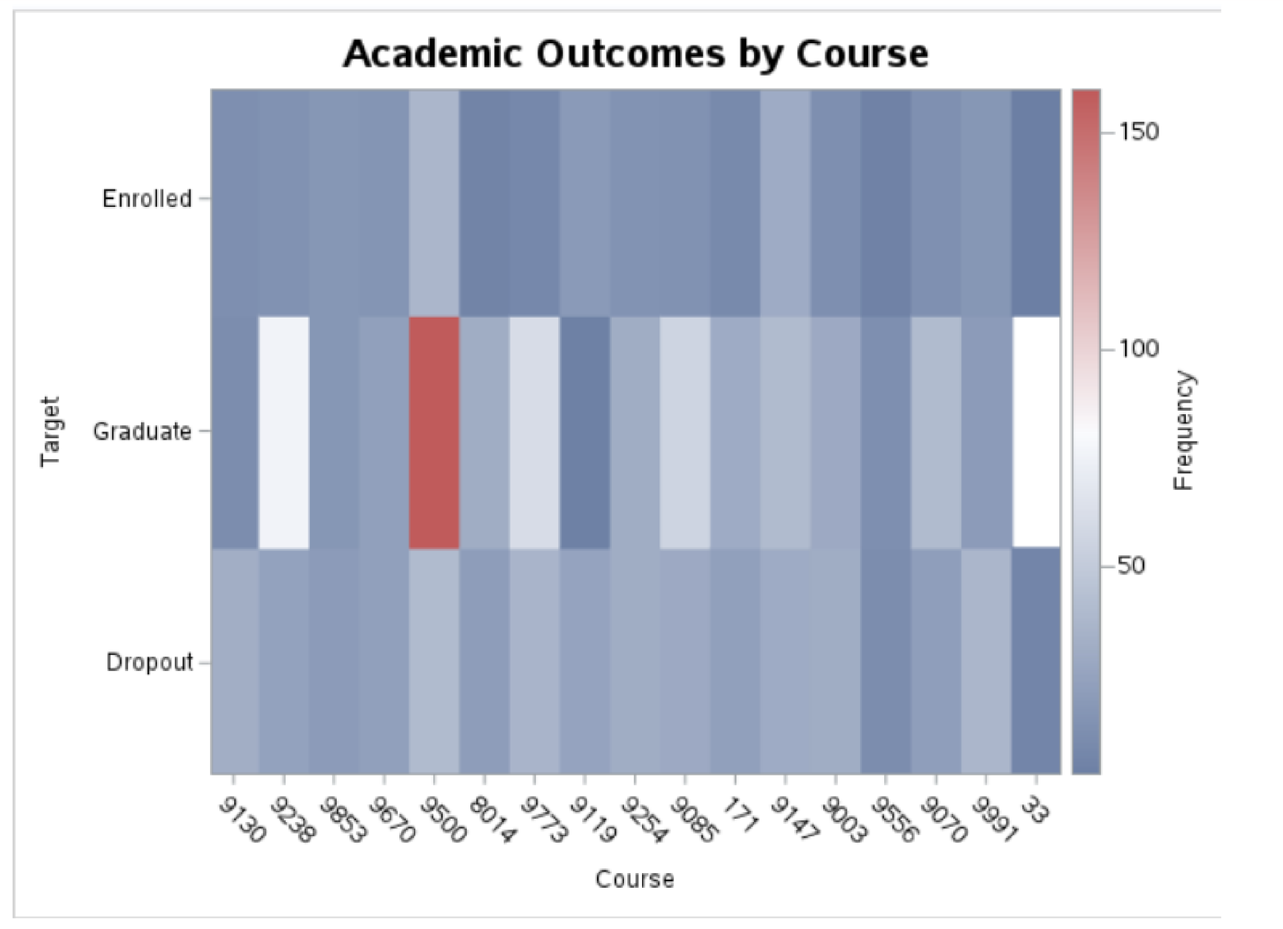
Correlation of Age At Enrollment With First & Second Semester Grade





Exploratory Data Analysis

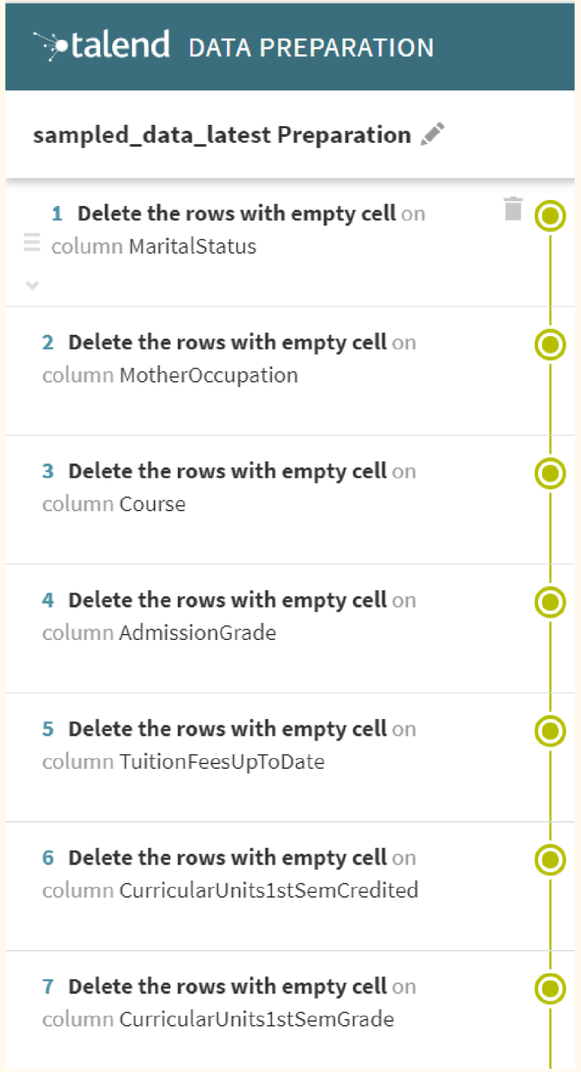
Academic Outcomes By Courses



MODIFY

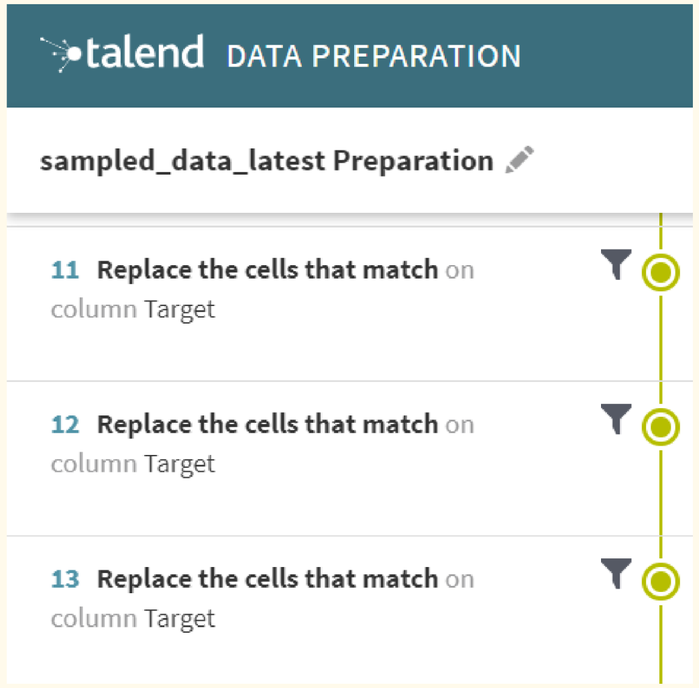
Data Preprocessing

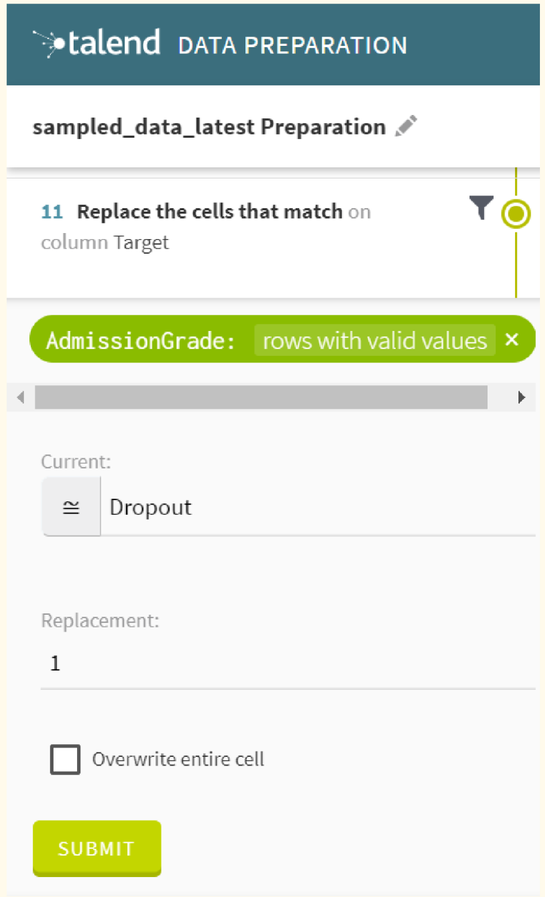
- Handle missing values

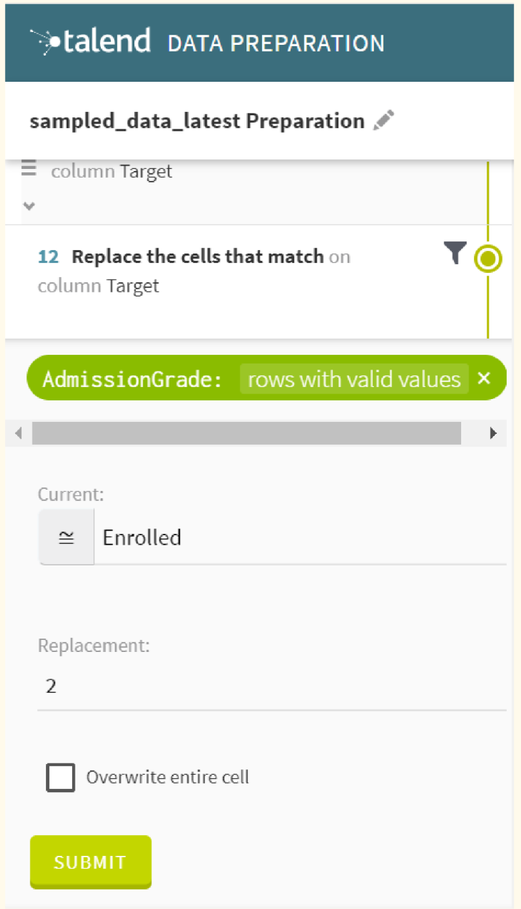


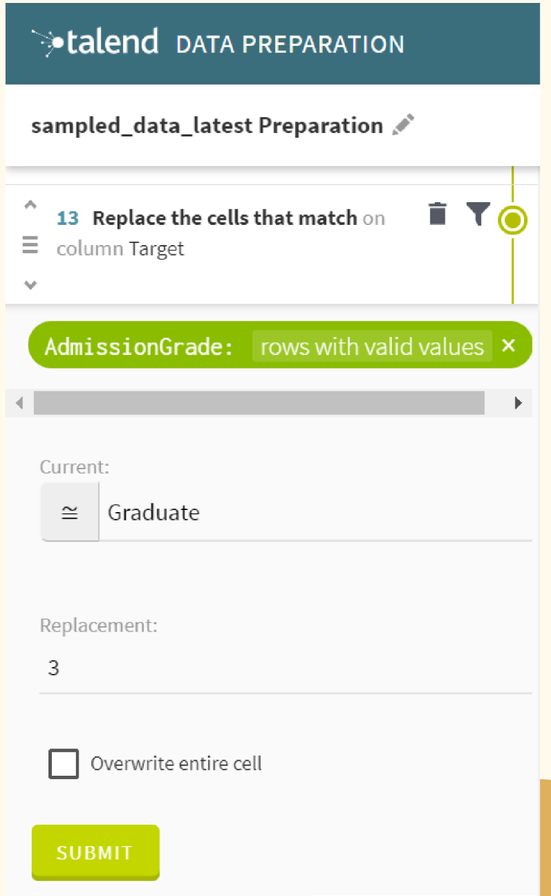
Data Preprocessing

- Encode categorical variables



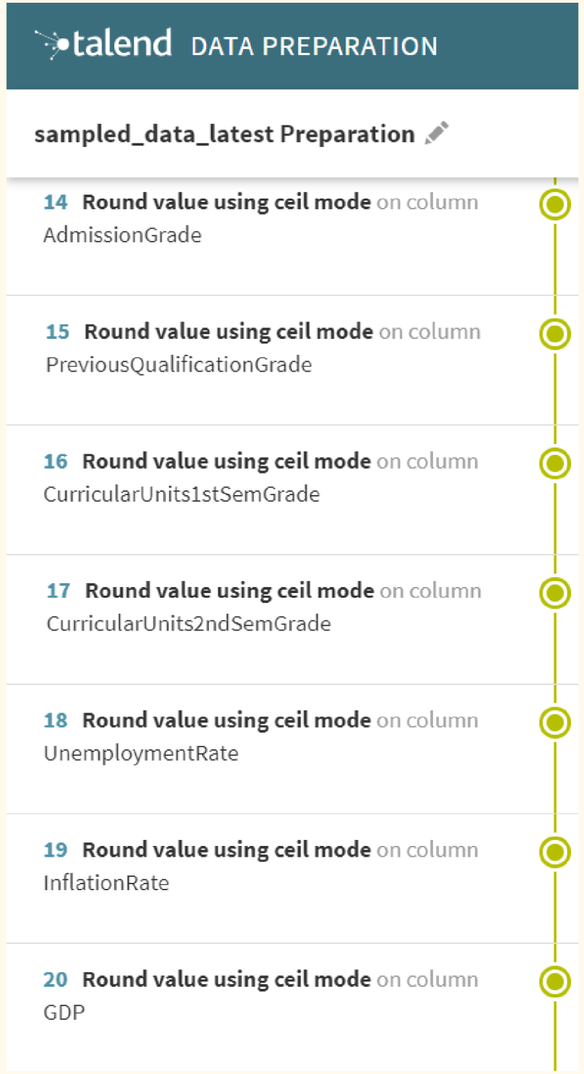






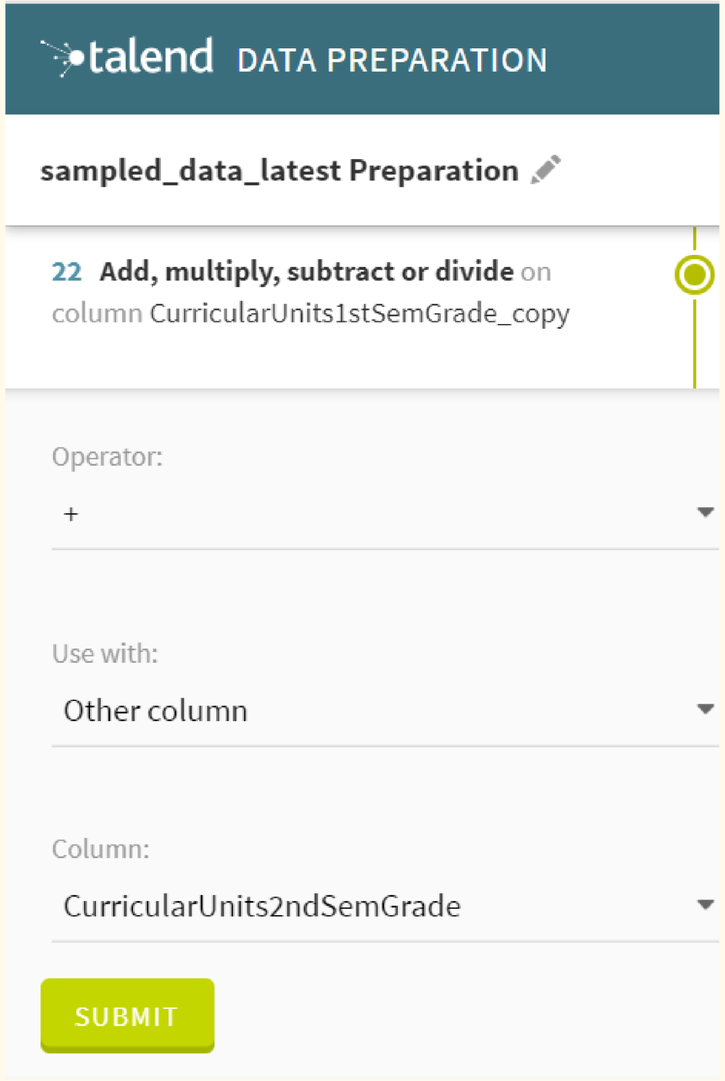
Data Preprocessing

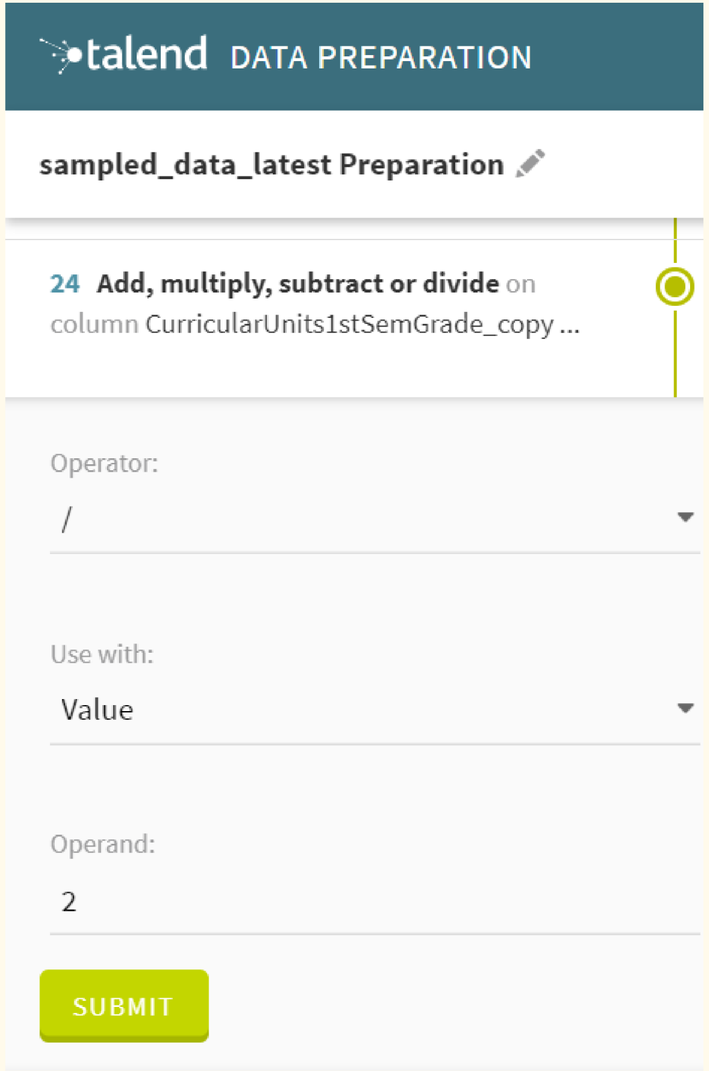
- Standardization of data



Data Preprocessing

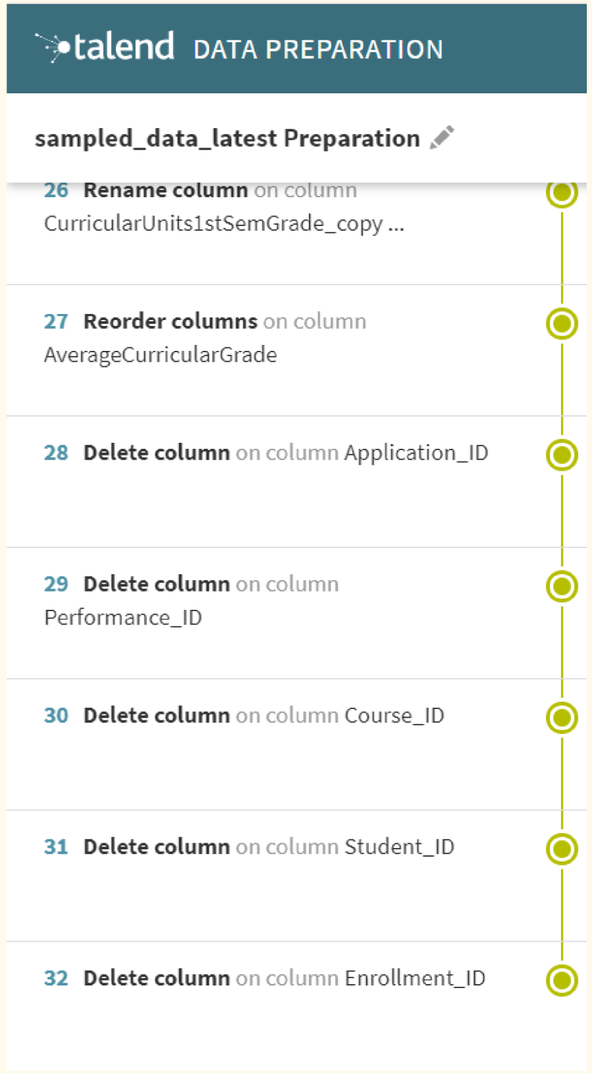
- Perform feature engineering





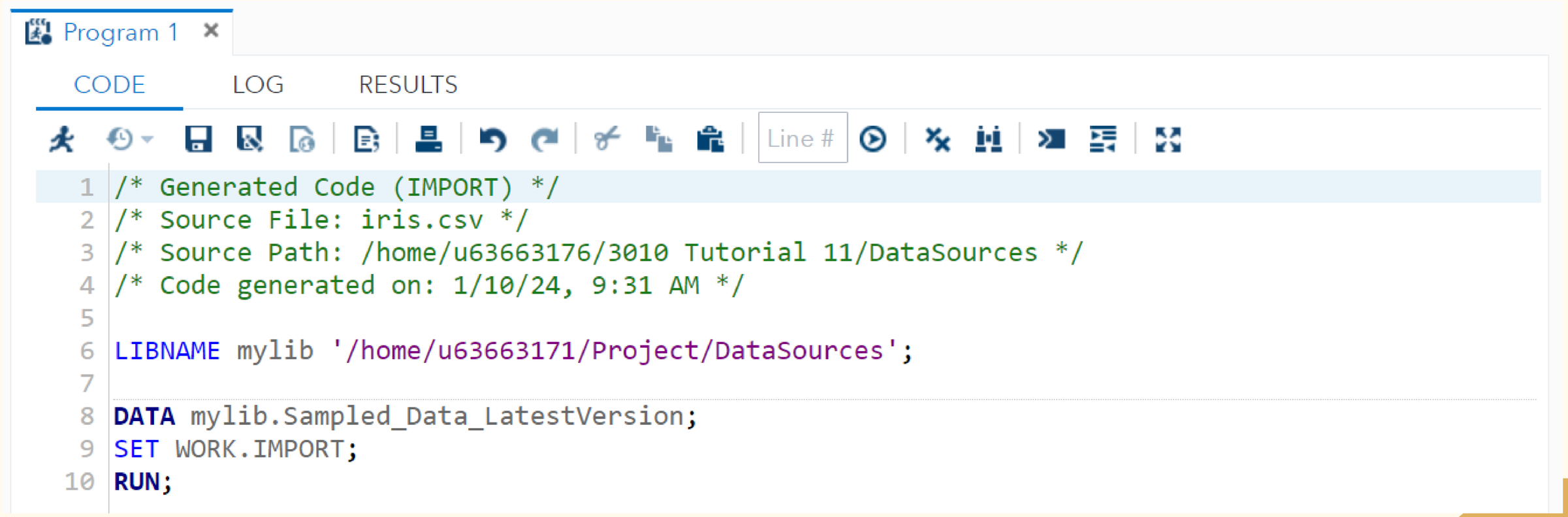
Data Preprocessing

- Remove unused identifiers



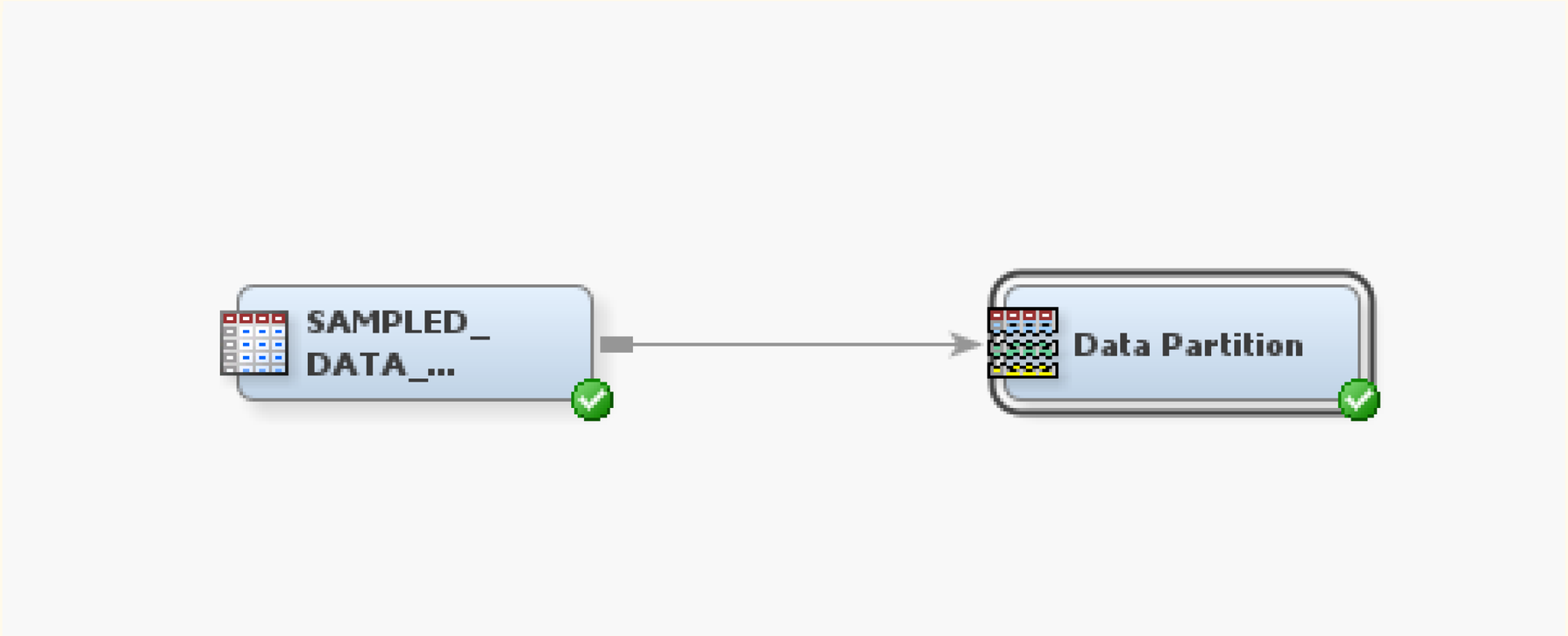
Data Preprocessing

- Code to convert the uploaded dataset to sas7bdat format



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

Rule Description

Map	Rule
RULE1	2 ==> 1
RULE2	3 ==> 1
RULE3	3 & 2 ==> 1
RULE4	1 ==> 2
RULE5	3 ==> 2
RULE6	3 & 1 ==> 2
RULE7	3 ==> 2 & 1
RULE8	2 ==> 3
RULE9	1 ==> 3
RULE10	2 & 1 ==> 3
RULE11	2 ==> 3 & 1
RULE12	1 ==> 3 & 2

Association Rule

Rule Table

Rules 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)

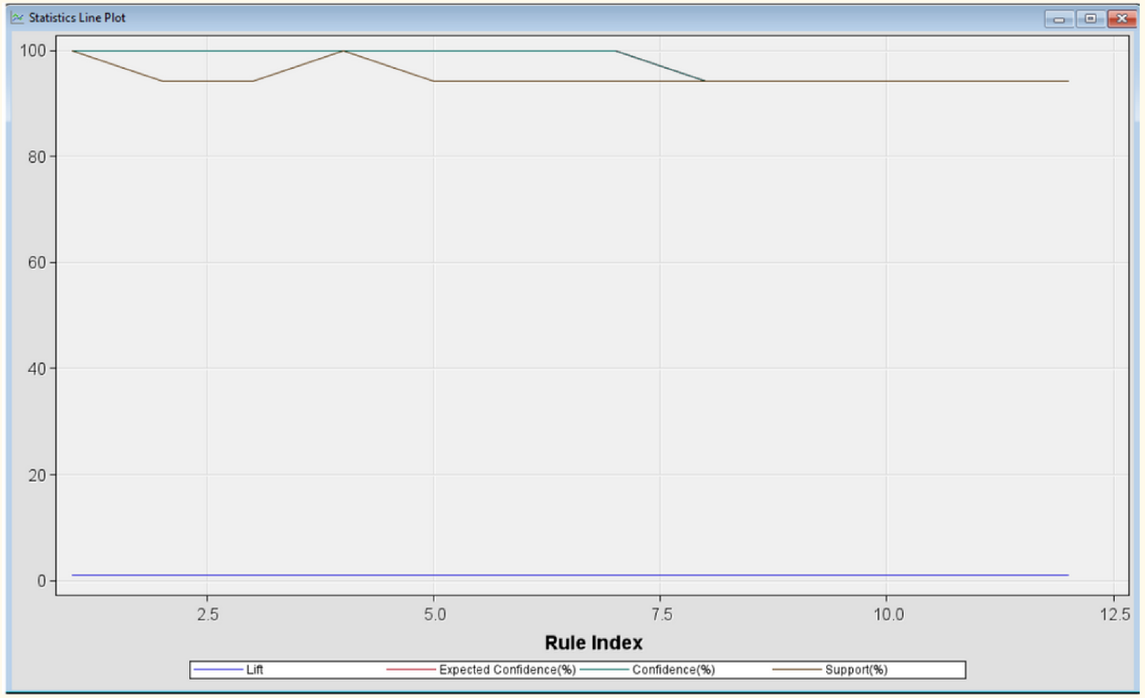
Results

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

Rule Statistics

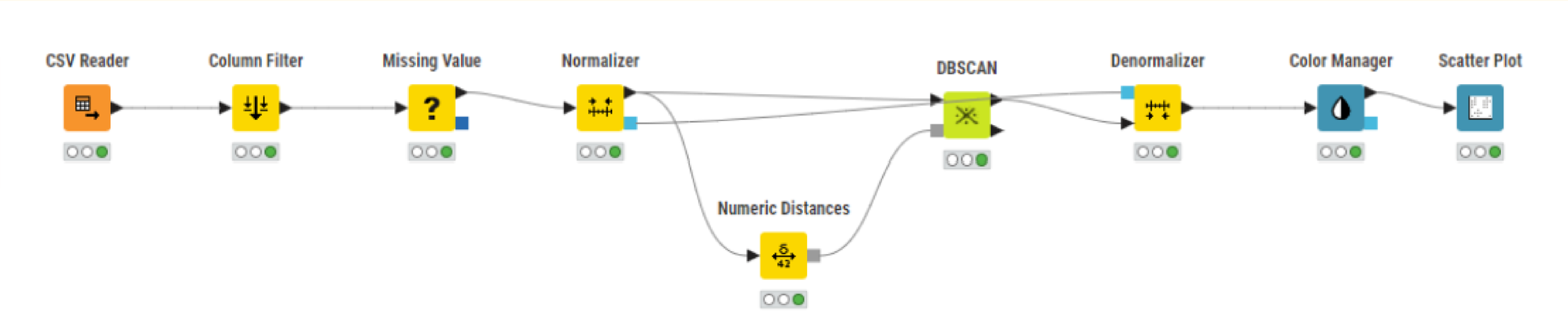
The MEANS Procedure

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



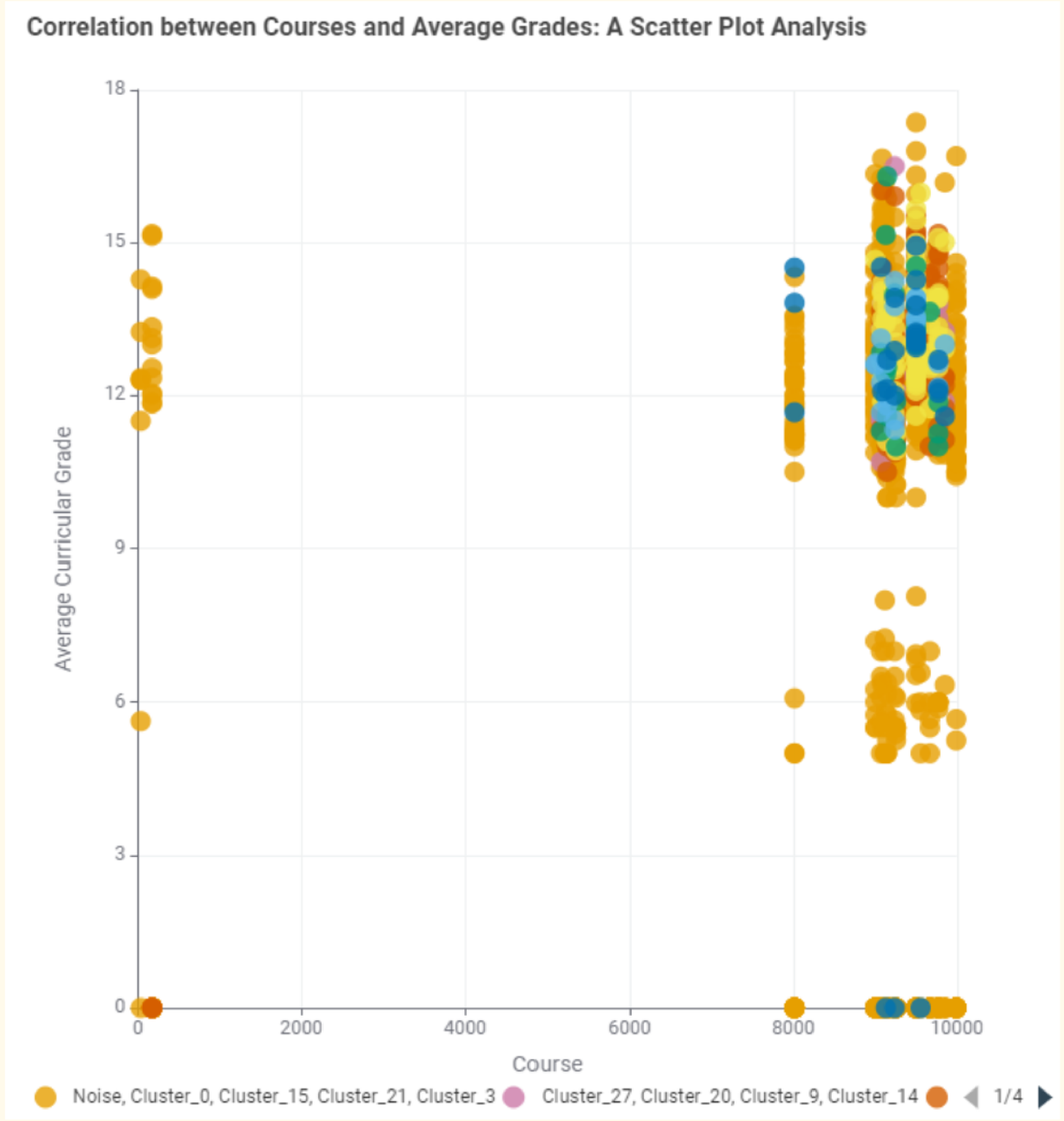
DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)



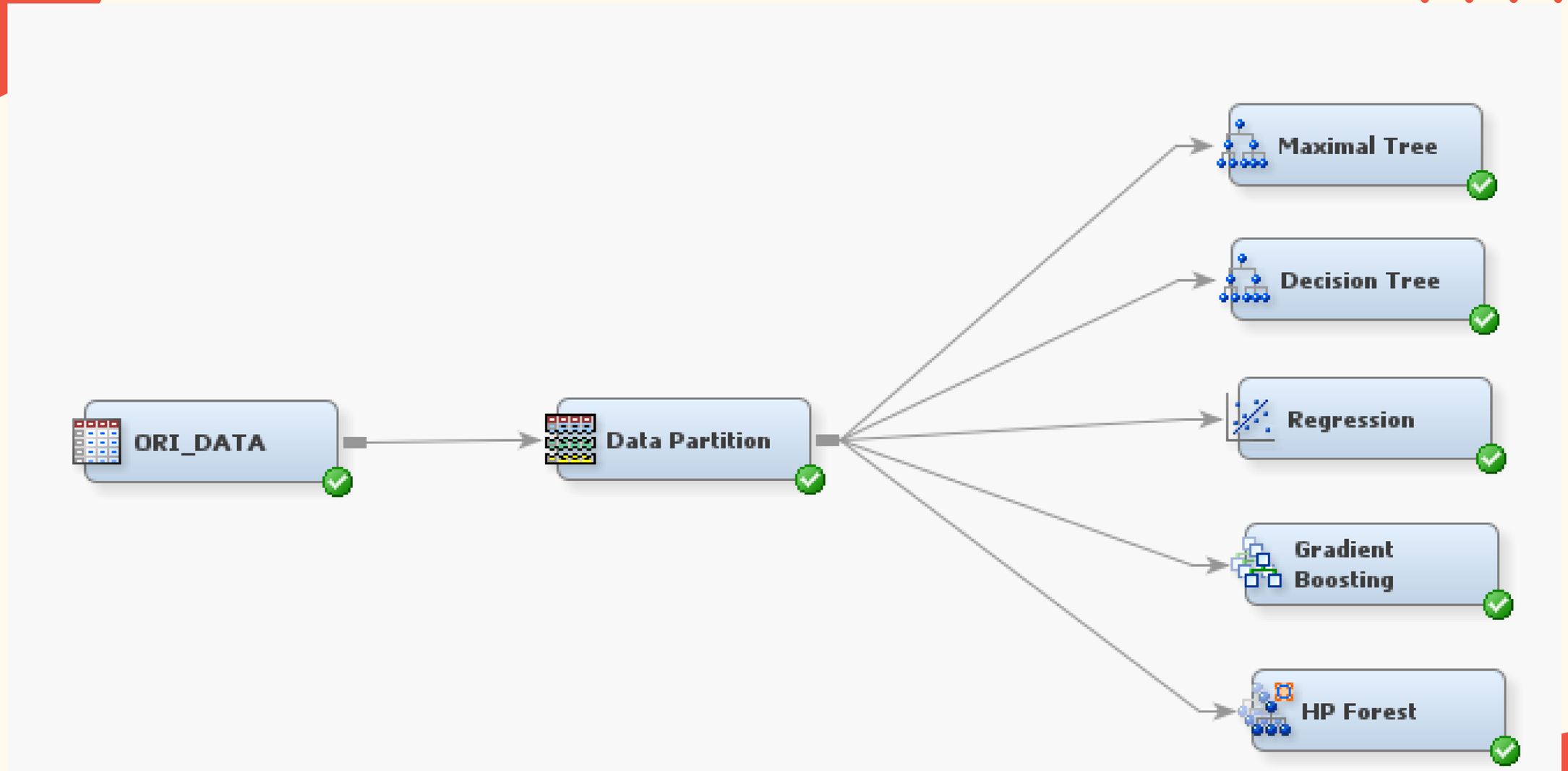
DBSCAN Diagram

DBSCAN



Scatter Plot Diagram

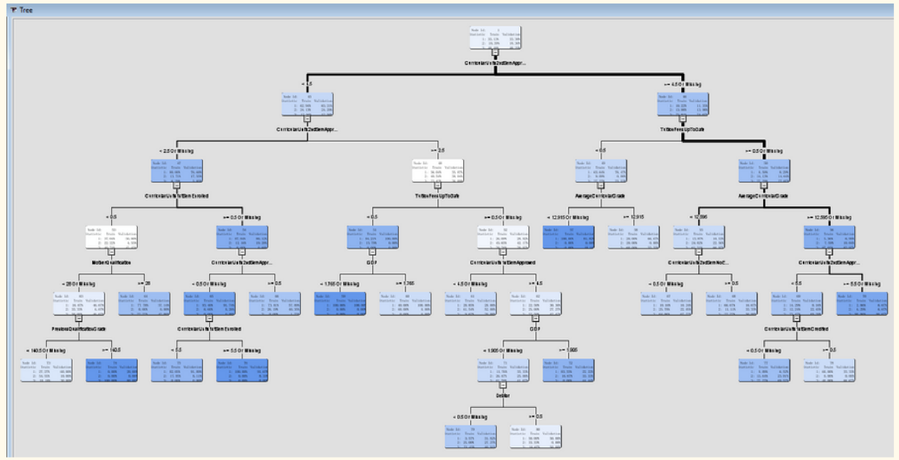
Model


ORI_DATA
Data Partition
Maximal Tree
Decision Tree
Regression
Gradient Boosting
HP Forest

Decision Tree (Maximal Tree)

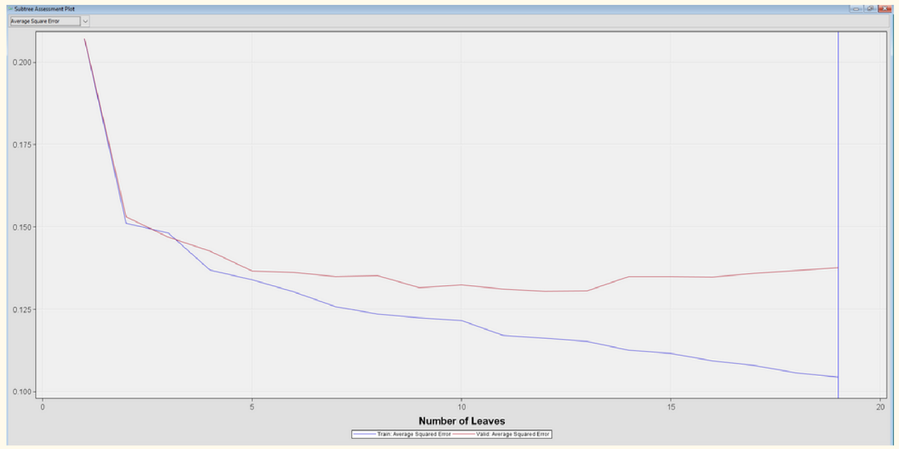
Tree

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



Misclassification Rate

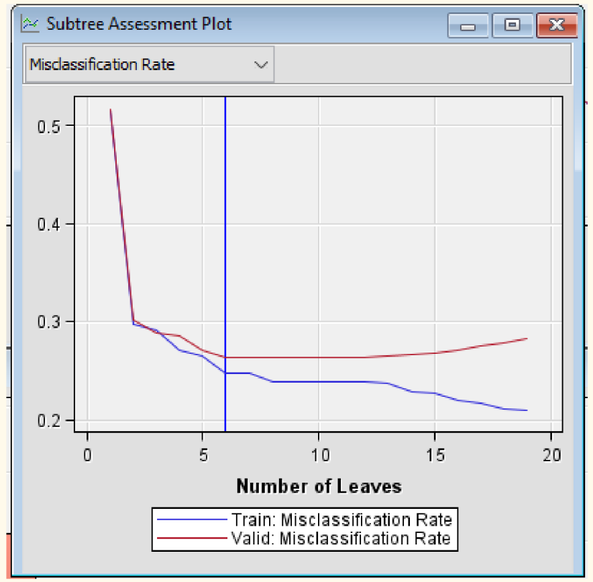
- Performance of validation sample improves up to a tree with approximately 5 leaves.
- Become worse as complexity increases.
- Overfitting - accurate predictions for test data but not valid data.



Decision Tree (Optimal Tree)

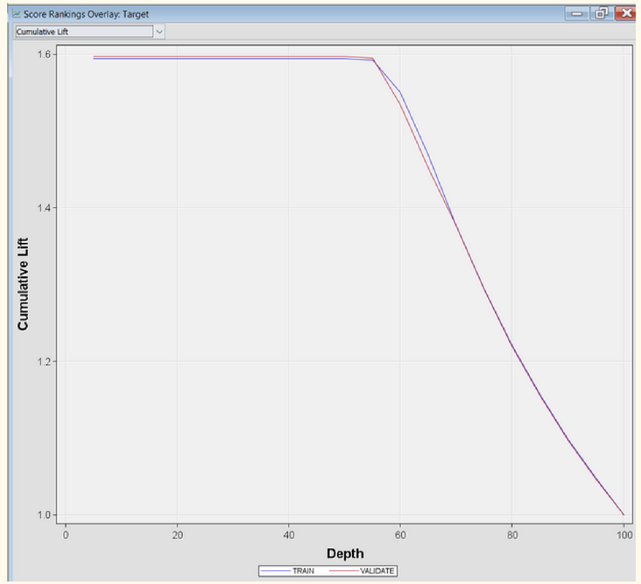
Misclassification Rate

- 6-leaf tree has lowest average squared error.



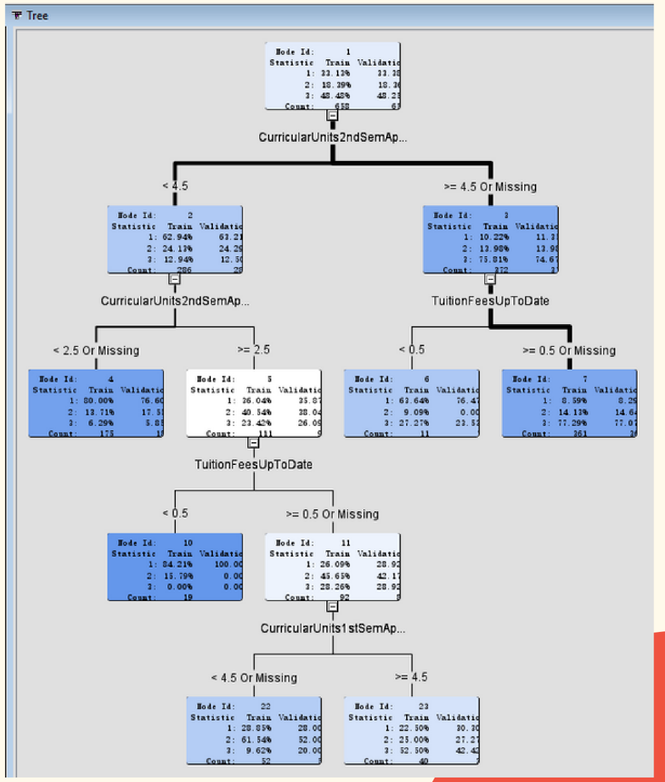
Cumulative Lift

- 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.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
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	2	9		5.4080	0.0669	846.9

Odds Ratio

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

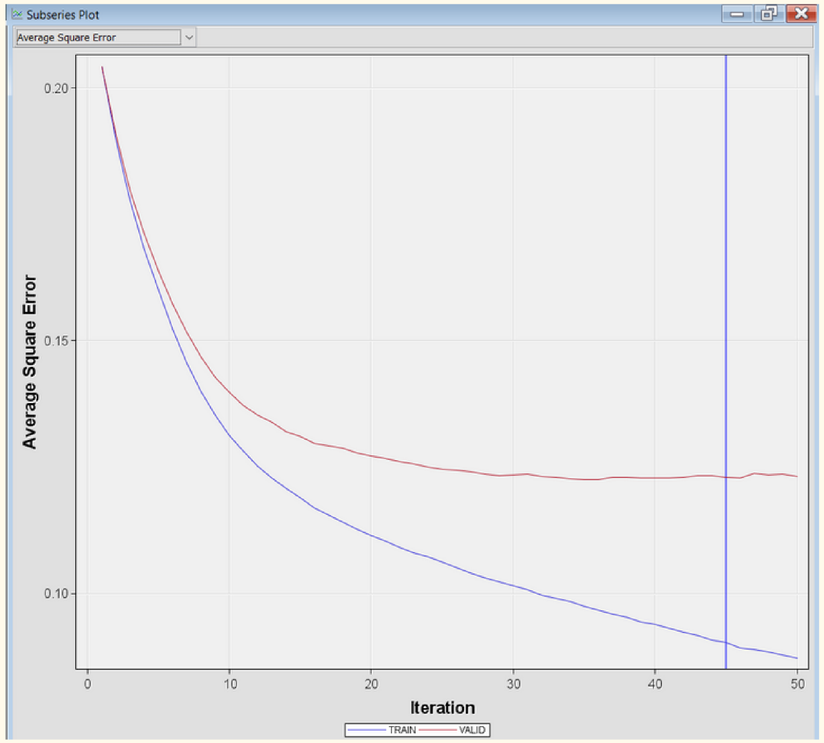
Odds Ratio Estimates

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

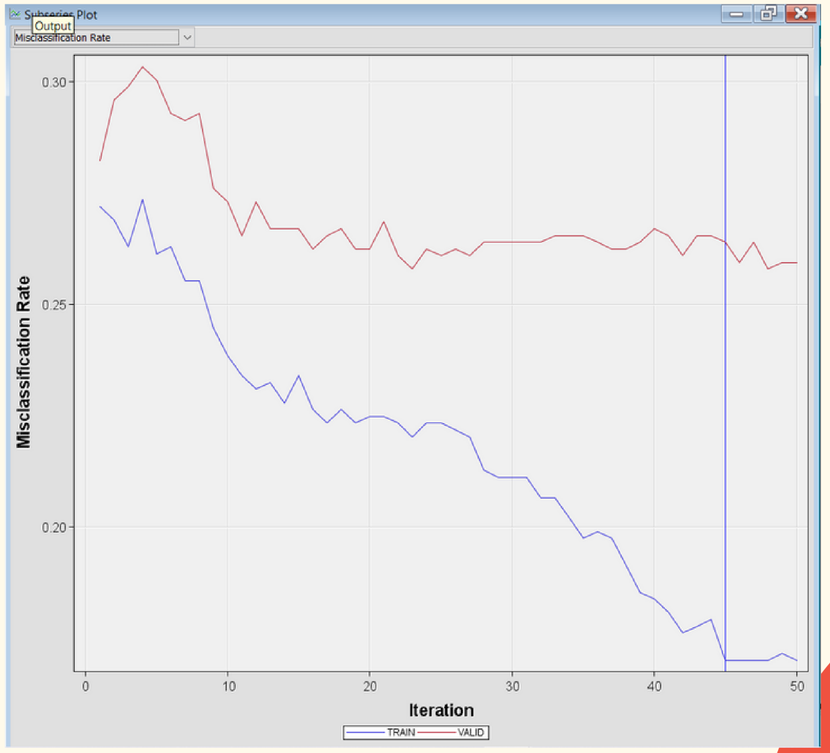
Average Square Error

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



Misclassification Rule

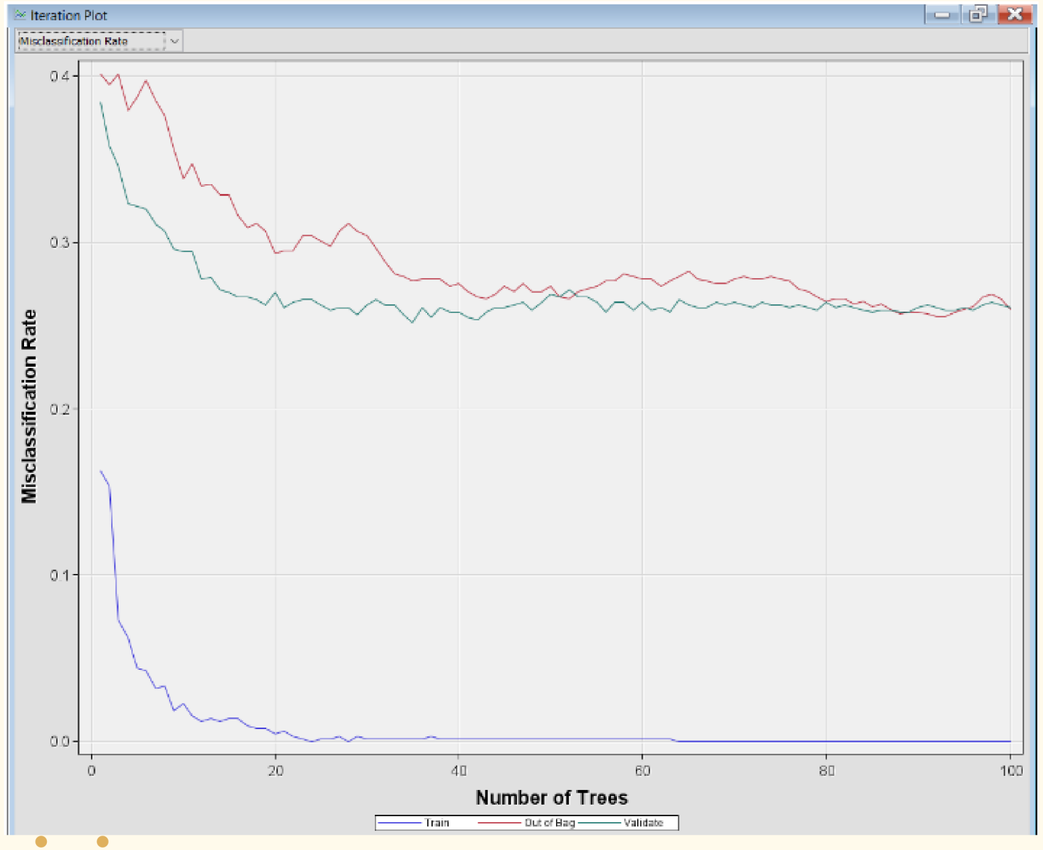
- When the training error decreases, the validation error increases.
- Find the optimal tree depth and stop the iteration before overfitting occurs



Random Forest

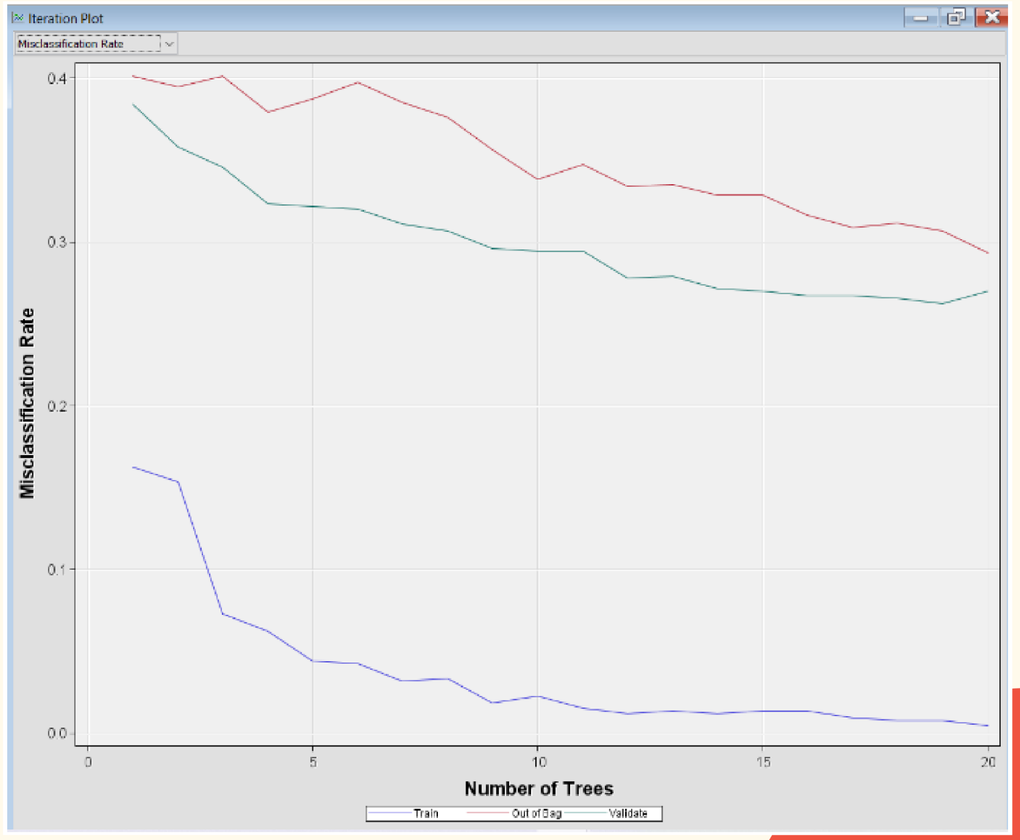
Before Reduction

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

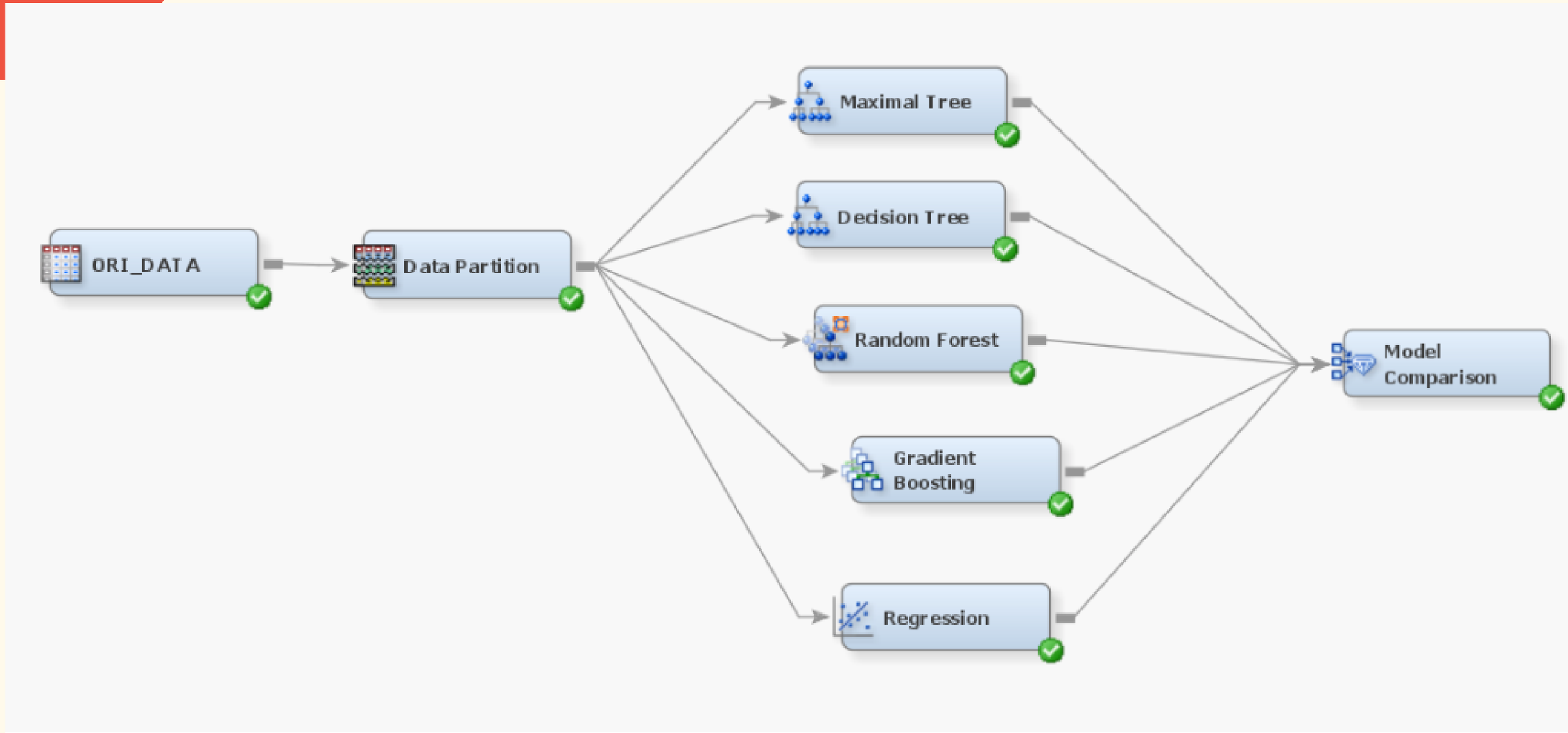


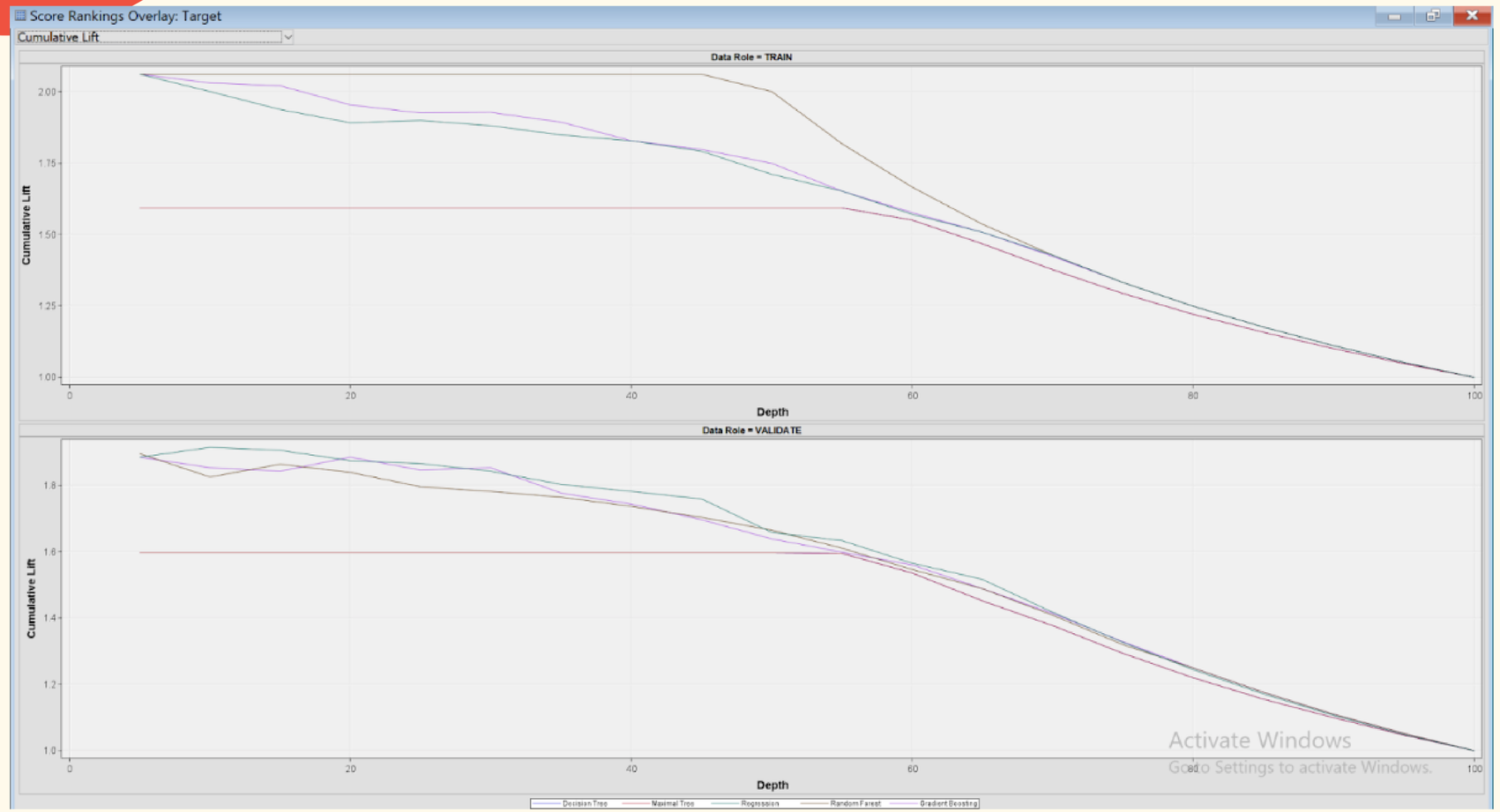
After Reduction

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



Assess


ORI_DATA
Data Partition
Maximal Tree
Decision Tree
Random Forest
Gradient Boosting
Regression
Model Comparison



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

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Squared Errors	Train: Sum of Case Weights Times Freq	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Average Error Function
Y	Reg	Reg	Regression	Target	Target	0.24	770	0.10	0.37	1	20	1	1	730	0.10	0.98	0.10	658	20	0.32	0.33	0.32	874	208	1	0.23	0.11	0.42
	Boost	Boost	Gradient Boosting	Target	Target	0.25		0.10				1	1			0.91		658		0.32				213	1	0.22	0.12	
	HPDMForest	HPDMForest	Random Forest	Target	Target	0.26		0.09					1			0.46		658		0.14				40.6	0		0.12	
	Tree	Tree	Maximal Tree	Target	Target	0.26		0.13				1	1			0.93		658		0.36				257		0.24	0.13	
	Tree2	Tree2	Decision Tree	Target	Target	0.26		0.13				1	1			0.93		658		0.36				257		0.24	0.13	

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