Yancie Troy H. Saludo 10/5/22

Assessment 8 - P3_0: Laboratory (Logistic Regression)

Context and Perspective

Yancie Troy Saludo is curious what factors plays a role in a person's graduate status, if they graduate or not, what significant data has relationship on their college degree as he is curious about his own academic future. He understands there are other factors that may influence the future of his academic progress so he went on to ask his peers to gain insight as he wants to investigate and analyze their data as he wonders what attributes have an effect in his own academic future.

For this activity, he will perform logistic regression to classify the relationship between graduated status of his peers and using various attributes to predict the if there is a significant factor on their graduated status. In doing so, he might have an idea on his own academic future.

Learning Objectives

After completing the activity, you should be able to:

- Explain what logistic regression is, how it is used and the benefits of using it.
- Recognize the necessary format for data in order to perform predictive logistic regression.
- Develop a logistic regression data mining model in Rapid Miner using a training data set.
- Interpret the model's outputs and get the accuracy score of the model and confusion matrix.

Organization Understanding

Yancie Troy's objective is to gain insight on the factors that may affect if the person has graduated from college so that it can aid him in his curiosity. He understands that there are several factors that does affect the academic future of a person. He has chosen logistic regression to get an estimate between the relationship of the graduated status on other attributes as he understands the benefits of it as it is easier to implement, interpret, and very efficient to train, effective for his classification objectives.

His data format also plays an important part since he is using binary data which is commonly used to perform predictive logistic regression. He will utilize the confusion matrix for further analysis and interpretation.

Data Understanding

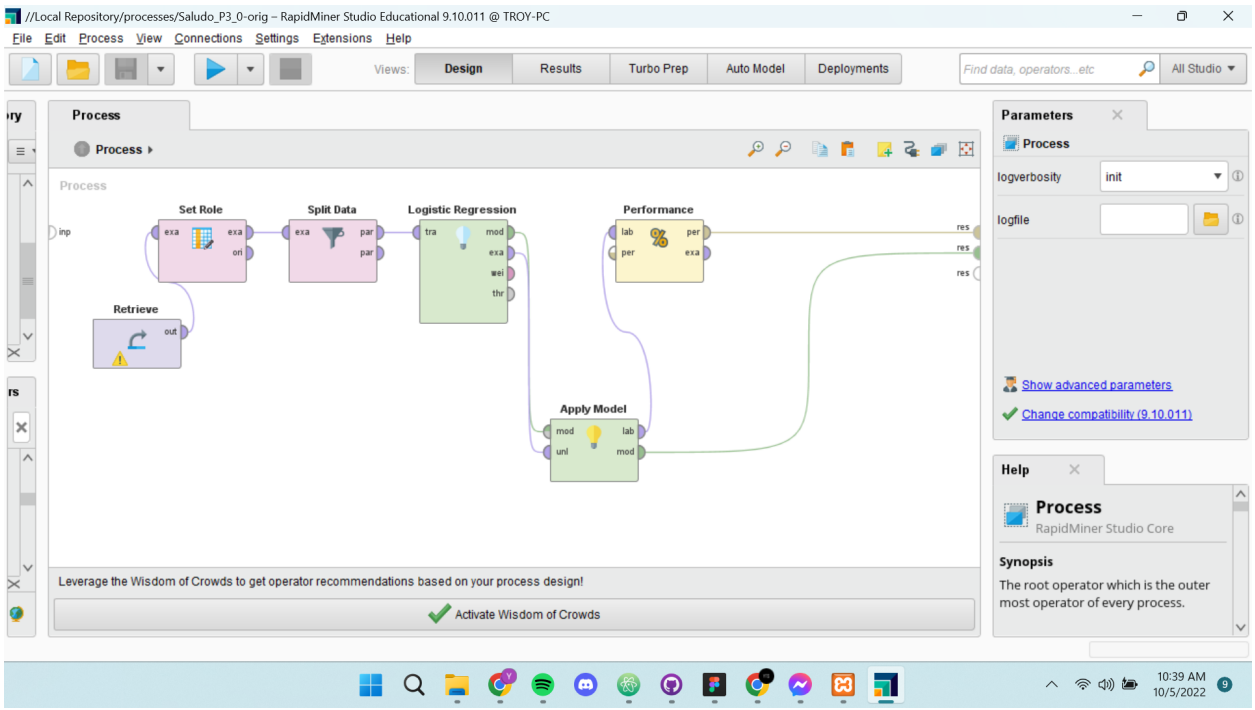
In his process of gaining insights, Troy has found 5 attributes and 100 observations that'll aid him in his development of the logistic regression model, he got it from for several adults that he knows who are at the age that they could have graduated from college by now. The dataset is made up of the following attributes:

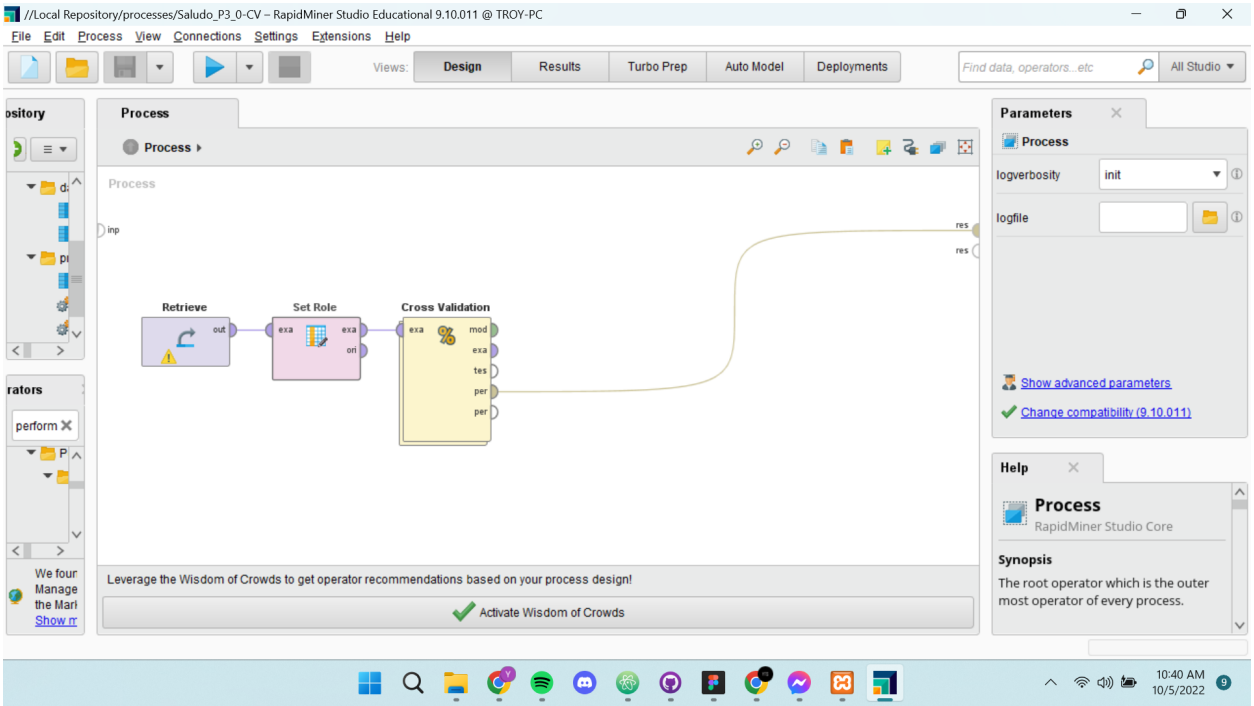
- Parent_Grad - 0 if neither of the person's parents graduated from college, 1 if one parent did, and a 2 if both parents did.
- Gender - 0 for female and 1 for male.
- Income_Level - 0 if the person lives in a household with a below average income, 1 for average, and 2 for above average income.
- Num_Siblings - number of siblings the person has.
- Graduated - 'Yes' if the person has graduated from college and 'No' if they have not.

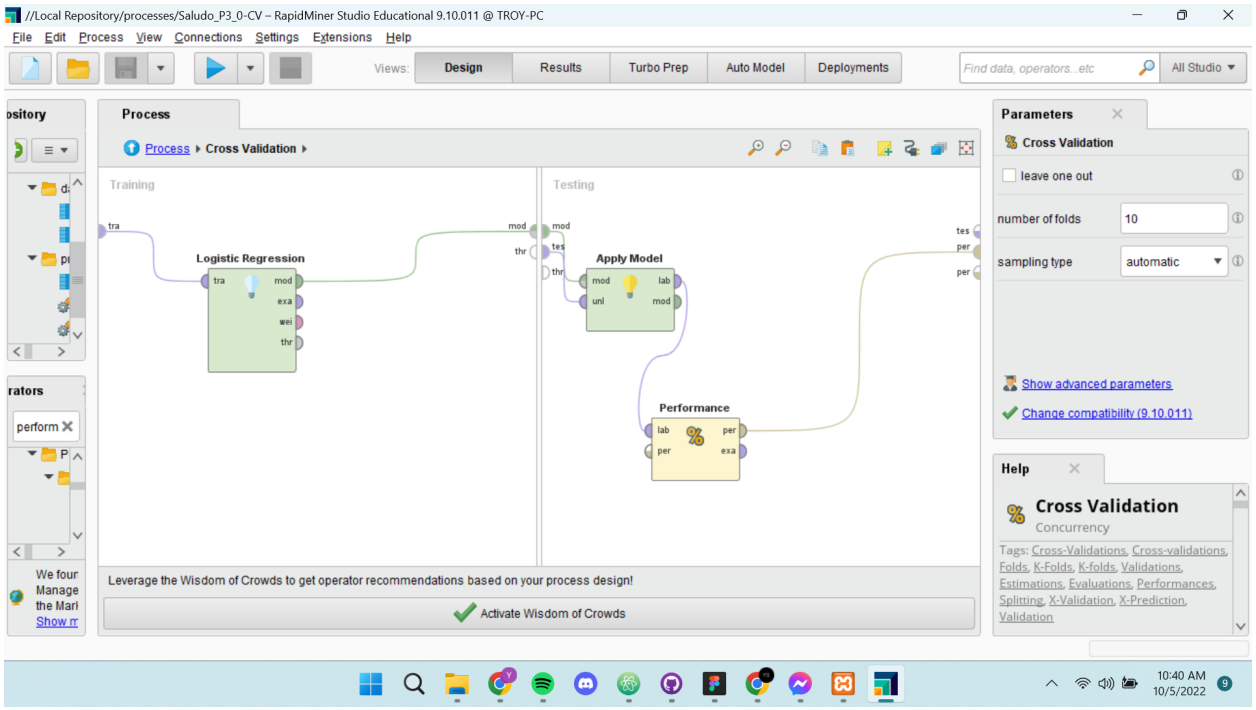
Data Preparation

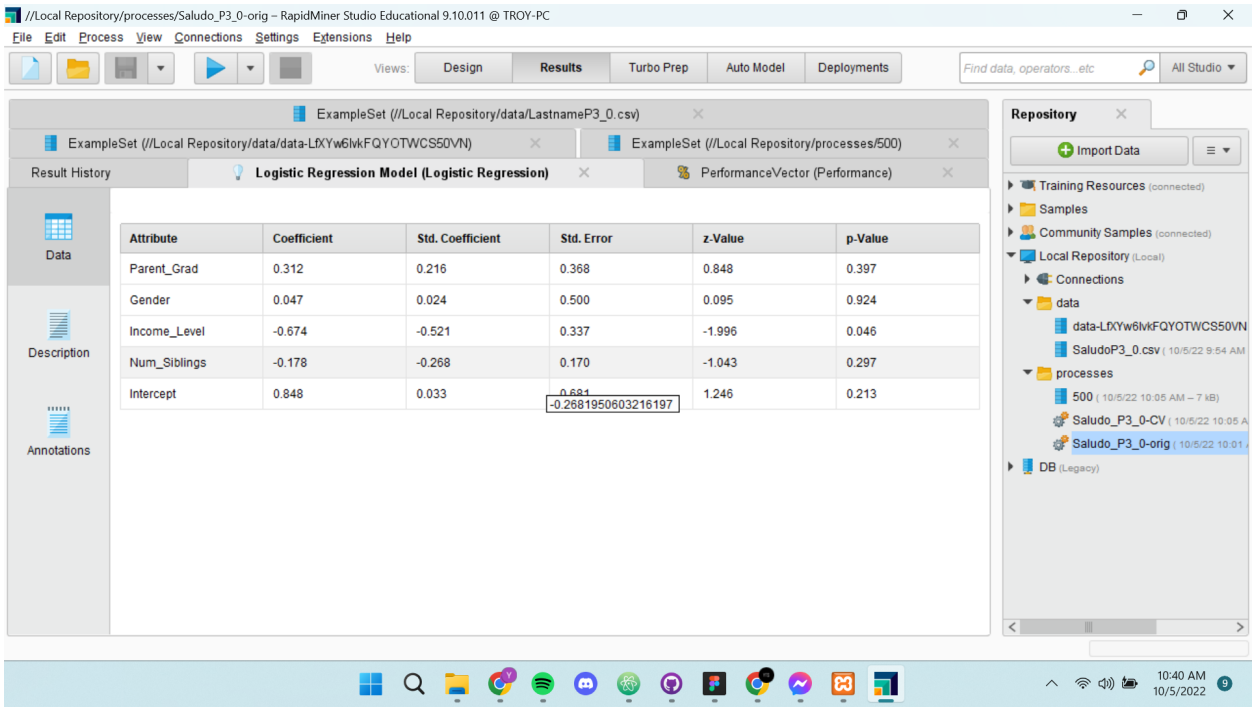
The CSV dataset taken from Troy's data gathering on several adults that he knows who are at the age that they could have graduated from college by now was then imported into RapidMiner data repository. It is a clean data so there was no need for data cleaning.

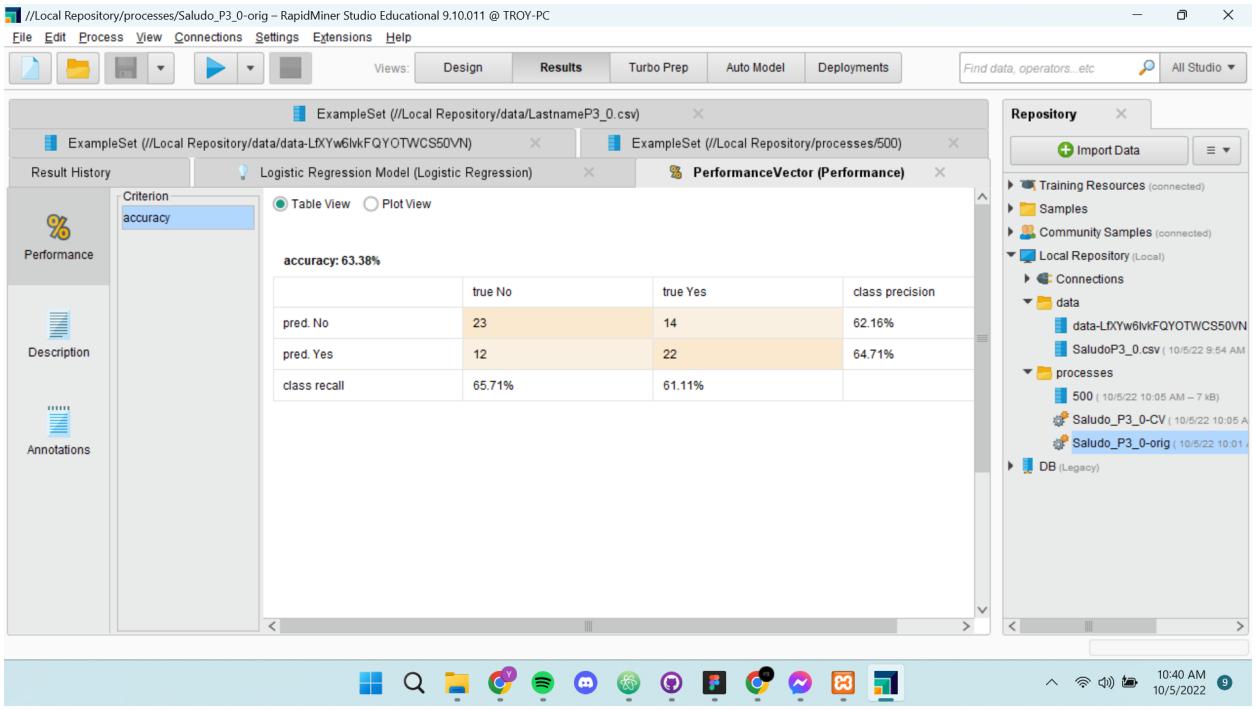
Data Modeling Demonstration

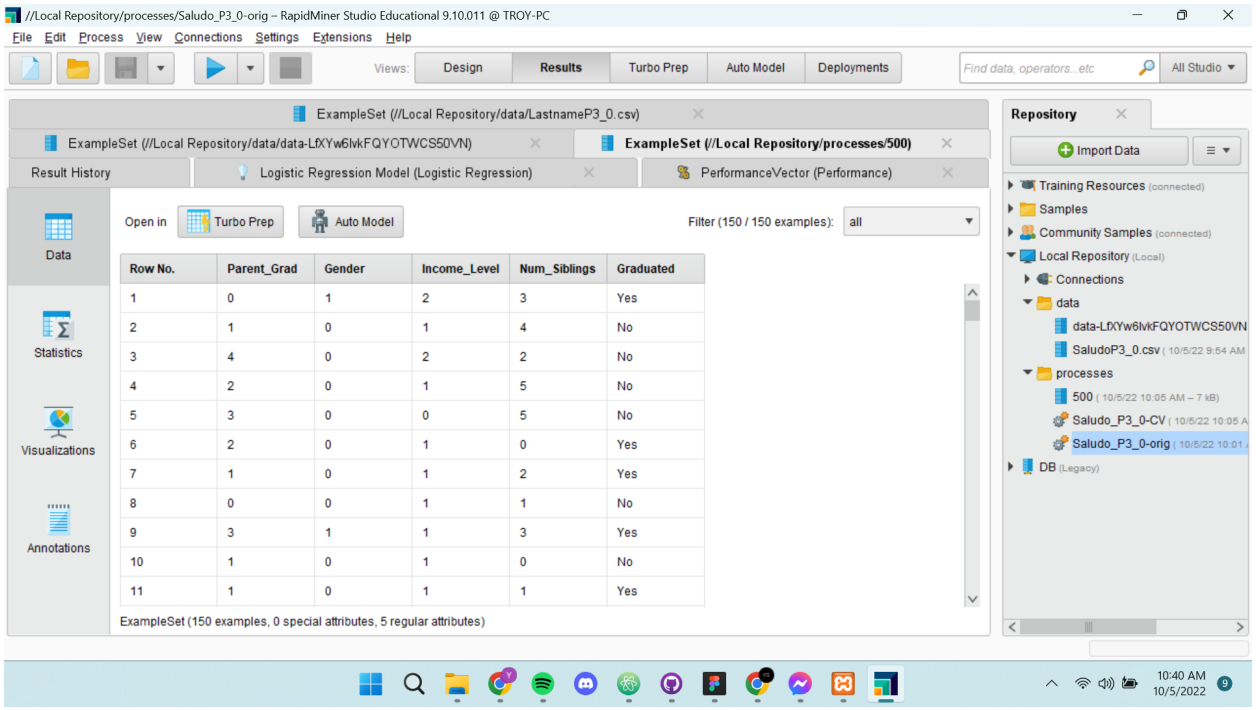


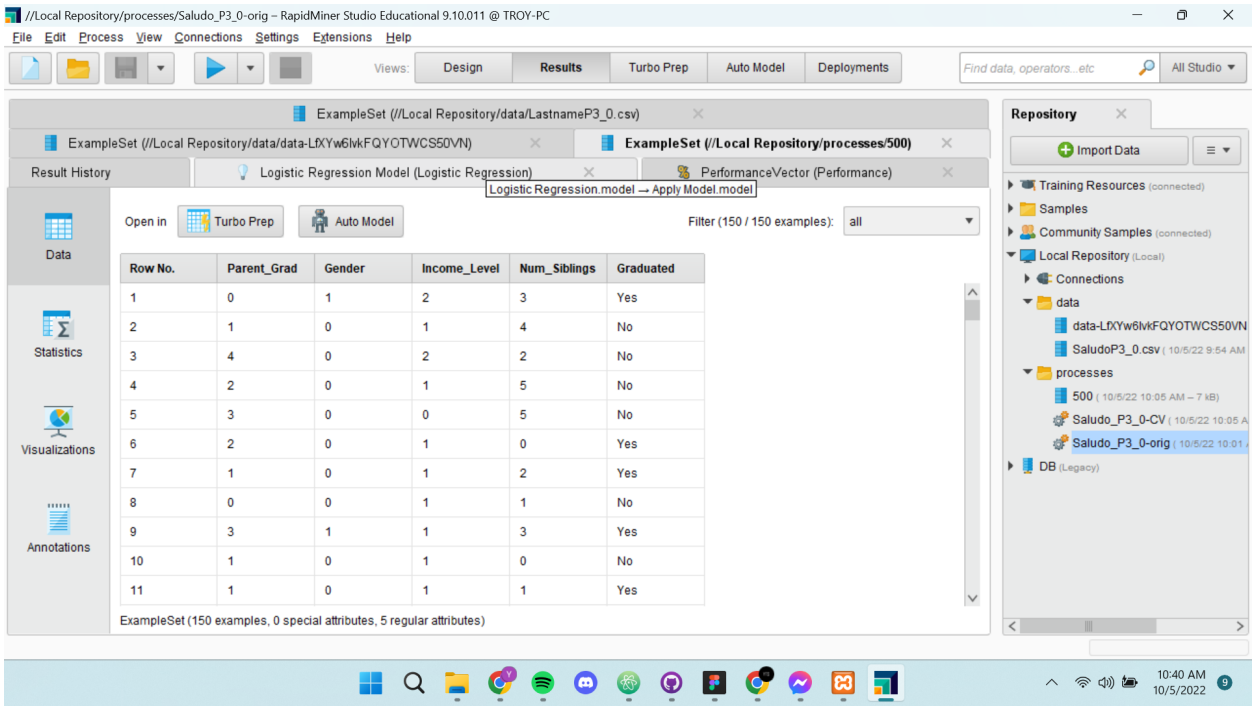


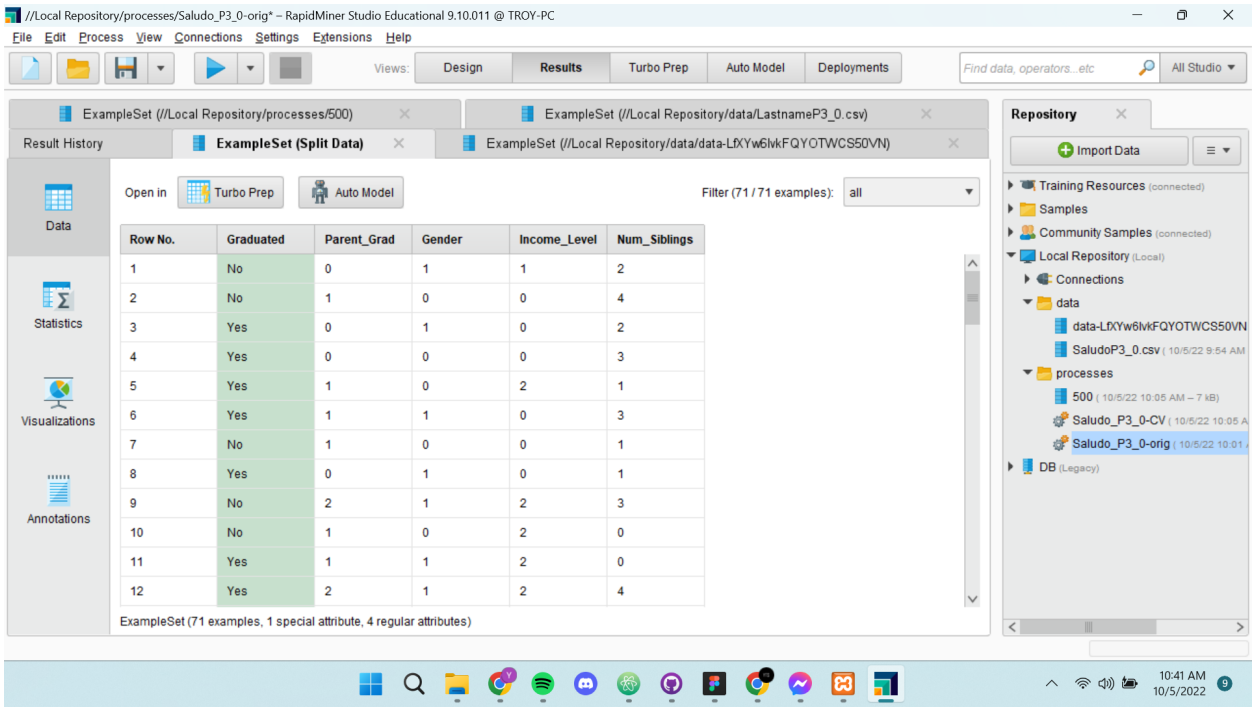

//Local Repository/processes/Saludo_P3_0-CV – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Process
Process ▸ Cross Validation ▸
Training
Testing
Logistic Regression
Apply Model
Performance
Parameters
Cross Validation
leave one out
number of folds
10
sampling type
automatic
Show advanced parameters
Change compatibility (9.10.011)
Help
Cross Validation
Concurrency
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
10:40 AM
10/5/2022

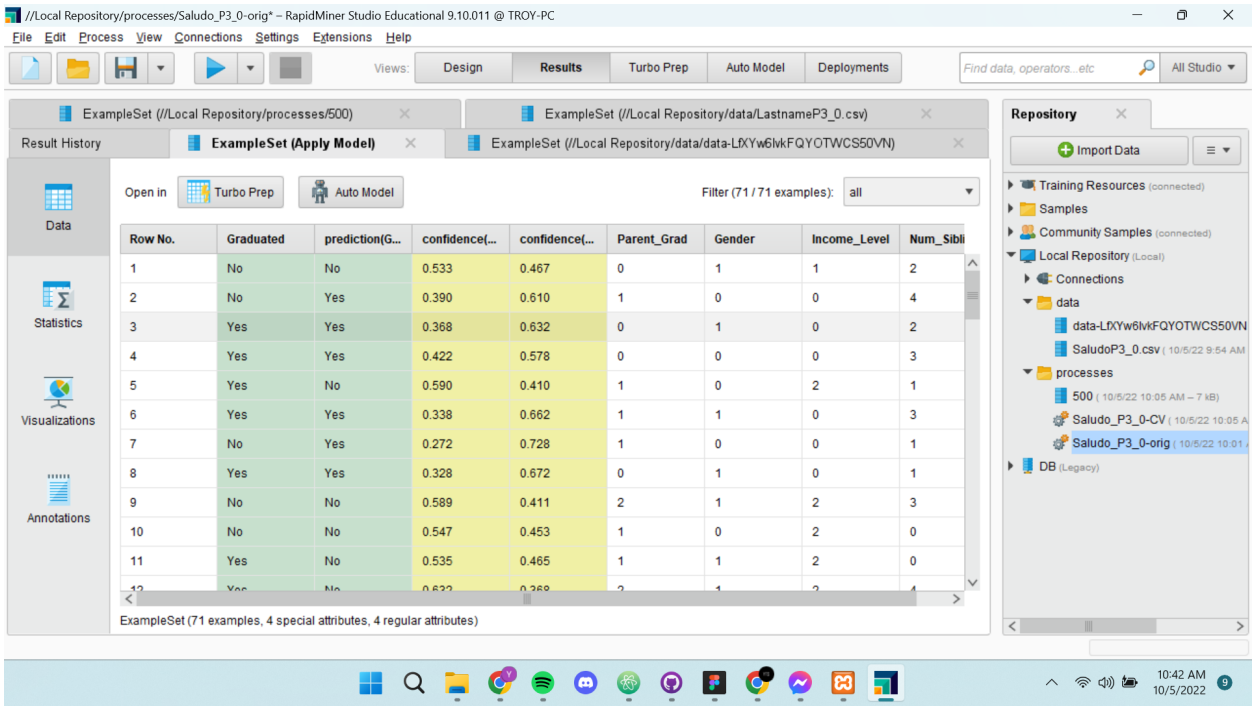

//Local Repository/processes/Saludo_P3_0-orig – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
ExampleSet (//Local Repository/processes/500)
Result History
Logistic Regression Model (Logistic Regression)
PerformanceVector (Performance)
Data
Description
Annotations
Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value
Parent_Grad | 0.312 | 0.216 | 0.368 | 0.848 | 0.397
Gender | 0.047 | 0.024 | 0.500 | 0.095 | 0.924
Income_Level | -0.674 | -0.521 | 0.337 | -1.996 | 0.046
Num_Siblings | -0.178 | -0.268 | 0.170 | -1.043 | 0.297
Intercept | 0.848 | 0.033 | -0.2681950603216197 | 1.246 | 0.213
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv
processes
500
Saludo_P3_0-CV
Saludo_P3_0-orig
DB (Legacy)
10:40 AM
10/5/2022

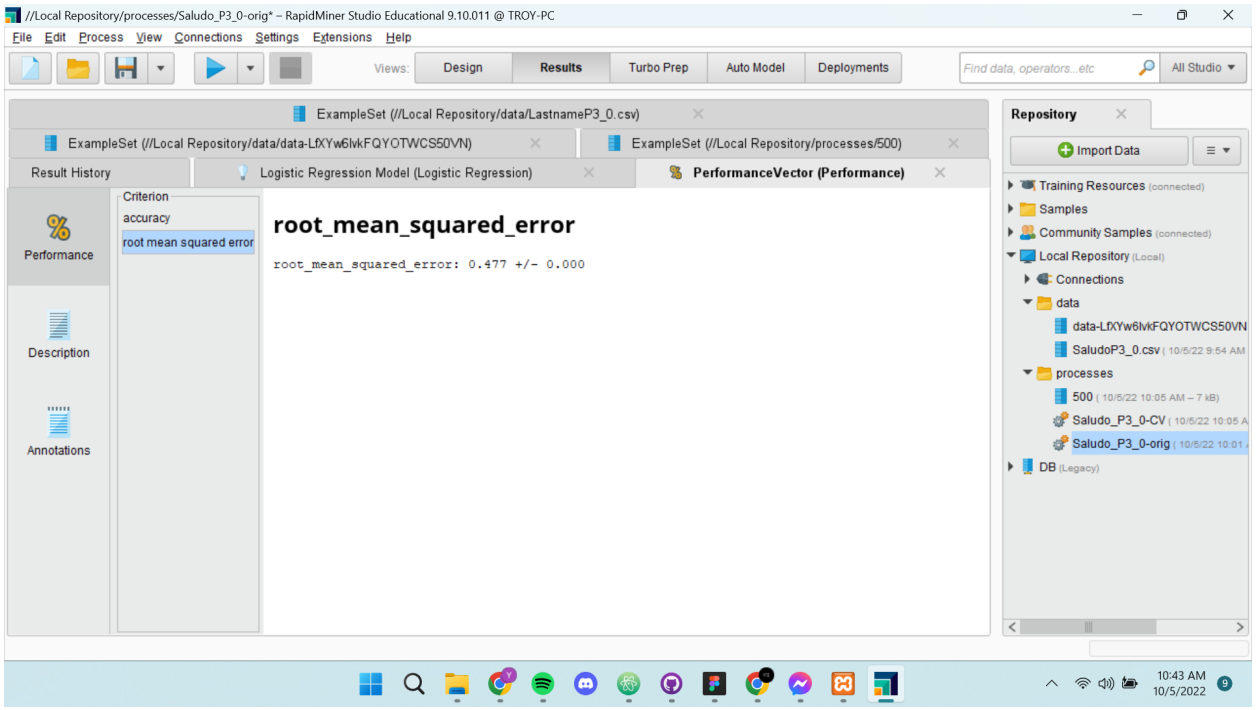

//Local Repository/processes/Saludo_P3_0-orig – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
ExampleSet (//Local Repository/processes/500)
Result History
Logistic Regression Model (Logistic Regression)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
accuracy
Table View
Plot View
accuracy: 63.38%
true No
true Yes
class precision
pred. No
23
14
62.16%
pred. Yes
12
22
64.71%
class recall
65.71%
61.11%
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv
processes
500
Saludo_P3_0-CV
Saludo_P3_0-orig
DB (Legacy)
10:40 AM
10/5/2022

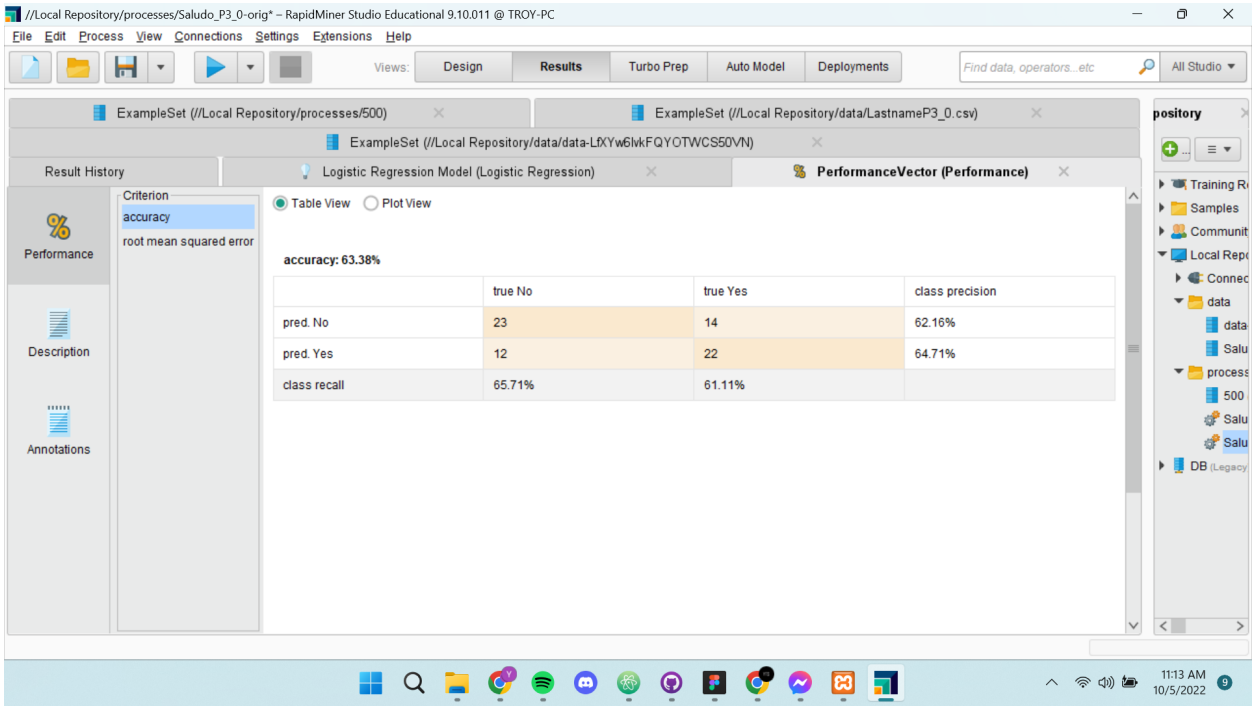

//Local Repository/processes/Saludo_P3_0-orig – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
ExampleSet (//Local Repository/processes/500)
Result History
Logistic Regression Model (Logistic Regression)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations
Open in
Turbo Prep
Auto Model
Filter (150 / 150 examples):
all
Row No.
Parent_Grad
Gender
Income_Level
Num_Siblings
Graduated
1 0 1 2 3 Yes
2 1 0 1 4 No
3 4 0 2 2 No
4 2 0 1 5 No
5 3 0 0 5 No
6 2 0 1 0 Yes
7 1 0 1 2 Yes
8 0 0 1 1 No
9 3 1 1 3 Yes
10 1 0 1 0 No
11 1 0 1 1 Yes
ExampleSet (150 examples, 0 special attributes, 5 regular attributes)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv
processes
500
Saludo_P3_0-CV
Saludo_P3_0-orig
DB (Legacy)
10:40 AM
10/5/2022


//Local Repository/processes/Saludo_P3_0-orig – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
ExampleSet (//Local Repository/processes/500)
Result History
Logistic Regression Model (Logistic Regression)
PerformanceVector (Performance)
Logistic Regression.model → Apply Model.model
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (150 / 150 examples): all
Row No. Parent_Grad Gender Income_Level Num_Siblings Graduated
1 0 1 2 3 Yes
2 1 0 1 4 No
3 4 0 2 2 No
4 2 0 1 5 No
5 3 0 0 5 No
6 2 0 1 0 Yes
7 1 0 1 2 Yes
8 0 0 1 1 No
9 3 1 1 3 Yes
10 1 0 1 0 No
11 1 0 1 1 Yes
ExampleSet (150 examples, 0 special attributes, 5 regular attributes)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv (10/5/22 9:54 AM
processes
500 (10/5/22 10:05 AM – 7 kB)
Saludo_P3_0-CV (10/5/22 10:05 A
Saludo_P3_0-orig (10/5/22 10:01
DB (Legacy)
10:40 AM
10/5/2022


//Local Repository/processes/Saludo_P3_0-orig* – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/processes/500)
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
Result History
ExampleSet (Split Data)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (71 / 71 examples): all
Row No. Graduated Parent_Grad Gender Income_Level Num_Siblings
1 No 0 1 1 2
2 No 1 0 0 4
3 Yes 0 1 0 2
4 Yes 0 0 0 3
5 Yes 1 0 2 1
6 Yes 1 1 0 3
7 No 1 0 0 1
8 Yes 0 1 0 1
9 No 2 1 2 3
10 No 1 0 2 0
11 Yes 1 1 2 0
12 Yes 2 1 2 4
ExampleSet (71 examples, 1 special attribute, 4 regular attributes)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv (10/5/22 9:54 AM
processes
500 (10/5/22 10:05 AM – 7 kB)
Saludo_P3_0-CV (10/5/22 10:05 A
Saludo_P3_0-orig (10/5/22 10:01
DB (Legacy)
10:41 AM
10/5/2022


//Local Repository/processes/Saludo_P3_0-orig* – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/processes/500)
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
Result History
ExampleSet (Apply Model)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (71 / 71 examples): all
Row No. Graduated prediction(G... confidence(... confidence(... Parent_Grad Gender Income_Level Num_Sibli
1 No No 0.533 0.467 0 1 1 2
2 No Yes 0.390 0.610 1 0 0 4
3 Yes Yes 0.368 0.632 0 1 0 2
4 Yes Yes 0.422 0.578 0 0 0 3
5 Yes No 0.590 0.410 1 0 2 1
6 Yes Yes 0.338 0.662 1 1 0 3
7 No Yes 0.272 0.728 1 0 0 1
8 Yes Yes 0.328 0.672 0 1 0 1
9 No No 0.589 0.411 2 1 2 3
10 No No 0.547 0.453 1 0 2 0
11 Yes No 0.535 0.465 1 1 2 0
ExampleSet (71 examples, 4 special attributes, 4 regular attributes)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv (10/5/22 9:54 AM
processes
500 (10/5/22 10:05 AM – 7 kB)
Saludo_P3_0-CV (10/5/22 10:05 A
Saludo_P3_0-orig (10/5/22 10:01
DB (Legacy)
10:42 AM
10/5/2022


//Local Repository/processes/Saludo_P3_0-orig* – RapidMiner Studio Educational 9.10.011 @ TROY-PC
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
ExampleSet (//Local Repository/data/LastnameP3_0.csv)
ExampleSet (//Local Repository/data/data-LfXYw6lvkFQYOTWCS50VN)
ExampleSet (//Local Repository/processes/500)
Result History
Logistic Regression Model (Logistic Regression)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
accuracy
root mean squared error
root_mean_squared_error
root_mean_squared_error: 0.477 +/- 0.000
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
data-LfXYw6lvkFQYOTWCS50VN
SaludoP3_0.csv (10/5/22 9:54 AM
processes
500 (10/5/22 10:05 AM – 7 kB)
Saludo_P3_0-CV (10/5/22 10:05 A
Saludo_P3_0-orig (10/5/22 10:01
DB (Legacy)
10:43 AM
10/5/2022



Evaluation and Deployment

In the evaluation and deployment phase, let's start with the understanding of what a confusion matrix is; used to determine the performance of the classification models for a given set of test data. In our case, we created a confusion matrix to assess the performance measurement for our machine learning classification. Through analyzation it is revealed that the model achieved a 63% accuracy with income value as the most significant attribute having a p-value of 0.046. It's not surprising that it is so as the same principles applies to the real world, income level does have an effect on completing their academic journey. Other variables included are num_siblings, and parent_grad which has some moderate significance.

With these interpretation parameters explained, we know head over to the confusion matrix interpretation:

accuracy: 63.38%

	true No	true Yes	class precision
pred. No	23	14	62.16%
pred. Yes	12	22	64.71%
class recall	65.71%	61.11%	

Starting with predicted no and the person really didn't graduate, we have 23 true positive values. And going to predicted yes and the person really graduated, we have 22 true negative values. Next up, did the model predicted that person didn't graduate but they did in fact graduate? We have 14 false positive values, and lastly, when the model predicted that the person graduated but they didn't in fact graduate, we have 12 false negative values.