AGENDA

1. Diagnostic Analytics with SPSS Modeler
2. Machine Learning using SPSS Modeler
3. Model Evaluation and Analysis using SPSS Modeler

Data Science Academy

I

Diagnostic Analytics with SPSS Modeler

Data Science Academy

Diagnostic analytics

Variables	Node	Description
Numeric variables	Regression Node	Provides correlation coefficients and significance levels between numerical variables.
1 categorical and numeric variables	Means Node	Compares the means between independent groups or between pairs of related fields to test whether a significant difference exists.
Categorical variables	CHAID Node	Tests for significance using a chi-square independence test
All variables	Feature Selection	Used to identify the fields that are most important for a given analysis.

Bütün hüquqlar qorunur.

Type Node

- Target variables are the output of the process. This is also called dependent or response variables.
- Predictor variables are the input for a process that is being analyzes. This is so called independent or explanatory variables.

Target variable

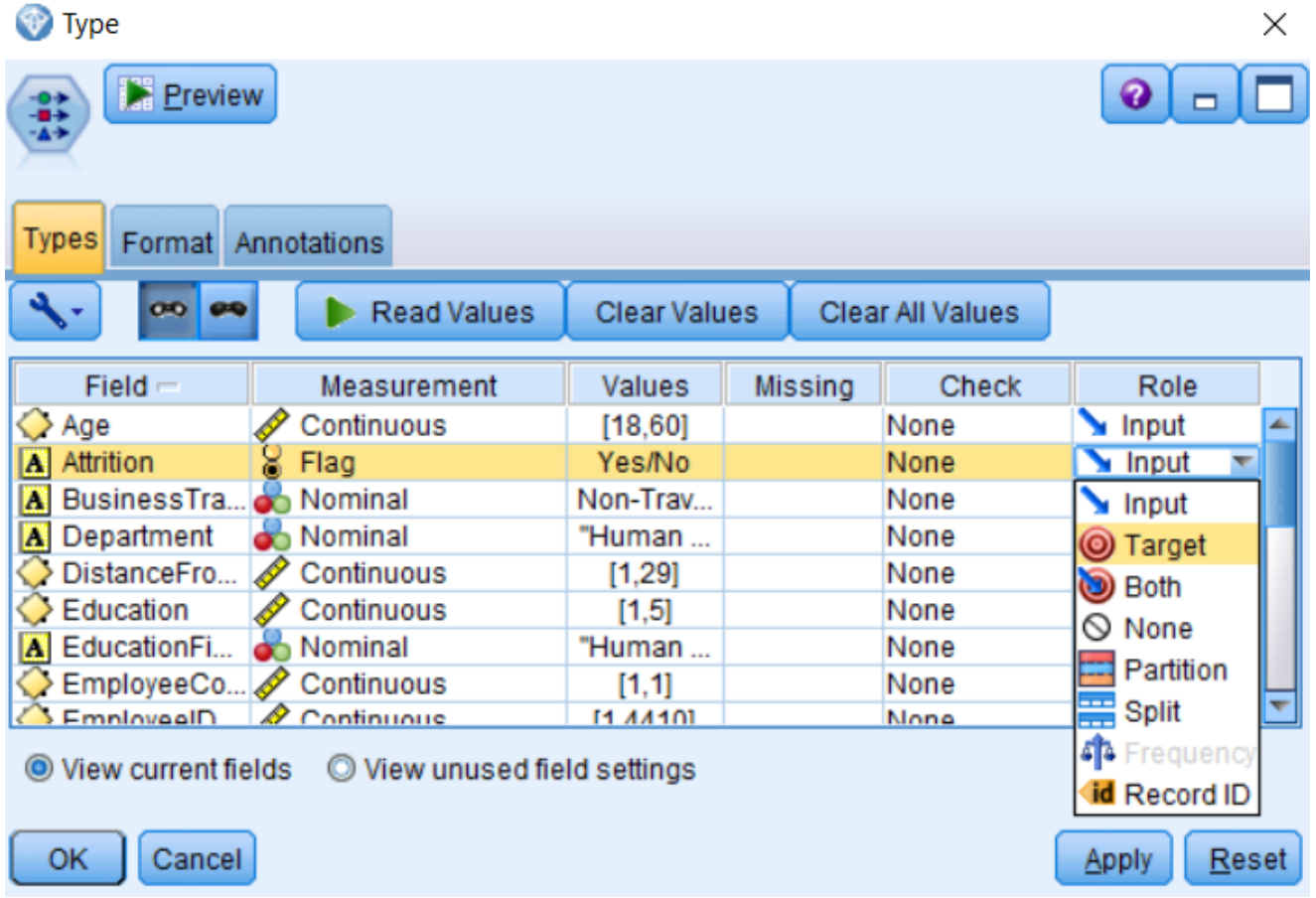
Input variables

Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	Gender	HourlyRate	JobLevel	MonthlyIncome
Yes	ravel_Rarely	1362.000	Resea	10.000	1.000	Medical	Male	32.000	1.000	1009.000
No	lon-Travel	287.000	Resea	5.000	2.000	Life Sciences	Male	73.000	1.000	1051.000
No	ravel_Rarely	1144.000	Sales	10.000	1.000	Medical	Male	74.000	1.000	1052.000
Yes	ravel_Rarely	945.000	Sales	9.000	3.000	Medical	Male	89.000	1.000	1081.000
Yes	ravel_Freque	746.000	Sales	24.000	3.000	Technical Deg	Male	45.000	1.000	1091.000
Yes	ravel_Rarely	303.000	Resea	2.000	3.000	Life Sciences	Male	47.000	1.000	1102.000
Yes	ravel_Freque	599.000	Sales	24.000	1.000	Life Sciences	Male	73.000	1.000	1118.000
No	ravel_Rarely	1276.000	Resea	2.000	1.000	Medical	Female	59.000	1.000	1129.000
No	ravel_Rarely	812.000	Sales	10.000	3.000	Medical	Female	69.000	1.000	1200.000
No	ravel_Rarely	373.000	Resea	1.000	2.000	Life Sciences	Male	47.000	1.000	1223.000
No	ravel_Rarely	391.000	Resea	15.000	2.000	Life Sciences	Male	96.000	1.000	1232.000
Yes	ravel_Rarely	202.000	Resea	8.000	3.000	Life Sciences	Female	34.000	1.000	1261.000
No	lon-Travel	990.000	Resea	7.000	3.000	Technical Deg	Male	64.000	1.000	1274.000
No	lon-Travel	727.000	Resea	3.000	3.000	Life Sciences	Male	41.000	1.000	1281.000
Yes	ravel_Freque	667.000	Sales	1.000	4.000	Life Sciences	Female	50.000	1.000	1359.000
Yes	ravel_Rarely	1259.000	Resea	2.000	4.000	Life Sciences	Male	95.000	1.000	1393.000
Yes	ravel_Rarely	1334.000	Resea	10.000	3.000	Life Sciences	Female	36.000	1.000	1416.000
Yes	ravel_Rarely	230.000	Resea	3.000	3.000	Life Sciences	Male	54.000	1.000	1420.000
No	ravel_Rarely	1181.000	Resea	3.000	1.000	Medical	Female	79.000	1.000	1483.000
No	lon-Travel	1431.000	Resea	14.000	3.000	Medical	Female	33.000	1.000	1514.000
Yes	ravel_Rarely	240.000	Human	22.000	1.000	Human Reso...	Male	58.000	1.000	1555.000
No	ravel_Rarely	1423.000	Resea	1.000	3.000	Life Sciences	Male	72.000	1.000	1563.000
Yes	ravel_Freque	544.000	Sales	3.000	2.000	Medical	Female	70.000	1.000	1569.000
Yes	ravel_Rarely	1243.000	Resea	6.000	3.000	Life Sciences	Male	63.000	1.000	1601.000

Bütün hüquqlar qorunur.

Selecting Target with Type Node

- Open ‘general_data.csv’
- Connect **Type Node**
- Double click the node
- Click **Read Values**
- Select Role of ‘Attrition’ as Target
- Click OK



Bütün hüquqlar qorunur.

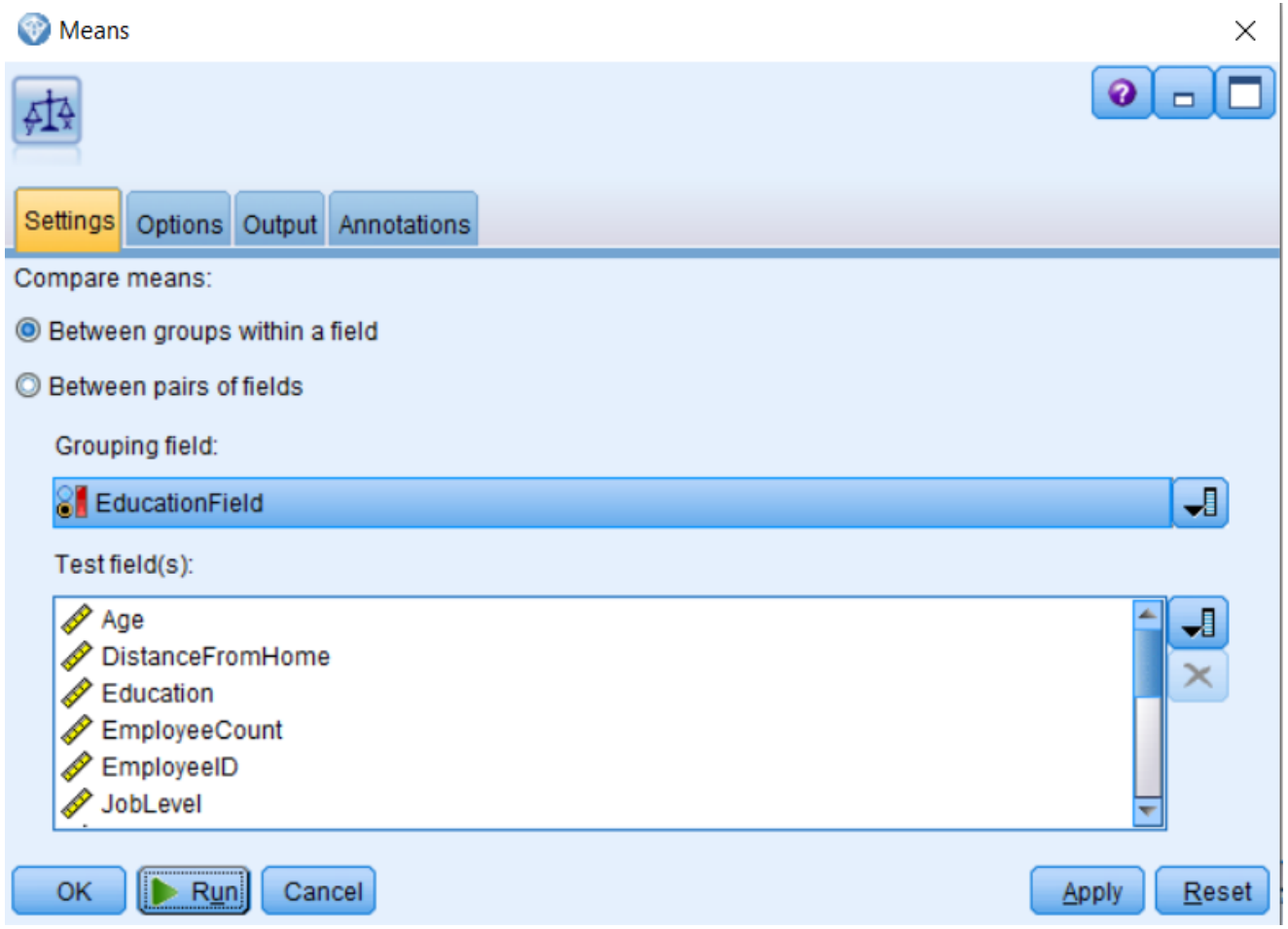
Means Node

- The Means node compares the means between independent groups or between pairs of related fields to test whether a significant difference exists. For example, you can compare mean revenues before and after running a promotion or compare revenues from customers who didn't receive the promotion with those who did.

Bütün hüquqlar qorunur.

Means Node

- Open ‘general_data.csv’
- Connect **Means Node**
- Double click the node
- Select ‘EducationField’ as Grouping Field
- Select all numeric variables in the Test Field(s)
- Run



Bütün hüquqlar qorunur.

Output

Means of [EducationField][Age DistanceFromHome Educatio...] #1

File Edit

Means Annotations

Sort by: Field View: Simple

Grouping field: EducationField

*Cells contain: Mean

Field	Life Scienc...	Other*	Medical*	Marketing*	Technical ...	Human Re...	Importance
Age	36.904	36.866	37.054	36.063	38.280	33.741	1.000 Important
DistanceFromHome	9.285	8.707	9.088	9.314	8.917	11.000	0.688 Unimportant
Education	2.921	3.110	2.881	2.887	2.924	2.778	0.969 Important
EmployeeCount	1.000	1.000	1.000	1.000	1.000	1.000	
EmployeeID	2205.041	2224.000	2199.922	2188.610	2213.538	2315.630	0.022 Unimportant
JobLevel	2.076	2.012	2.013	2.176	2.068	2.148	0.894 Unimportant
MonthlyIncome	65311.832	65761.463	67096.724	59030.881	64171.667	60452.963	0.957 Important
NumCompaniesWorked	2.717	2.672	2.648	2.785	2.809	1.975	0.880 Unimportant
PercentSalaryHike	15.254	15.500	15.366	14.918	14.712	14.778	0.993 Important
StandardHours	8.000	8.000	8.000	8.000	8.000	8.000	
StockOptionLevel	0.751	0.817	0.873	0.774	0.720	0.815	0.999 Important
TrainingTimesLastYear	2.850	2.634	2.780	2.792	2.795	2.556	0.926

Bütün hüquqlar qorunur.

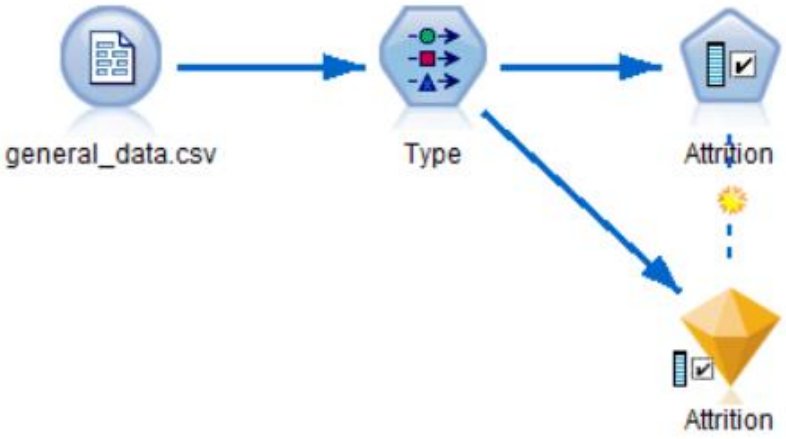
Feature Selection Node

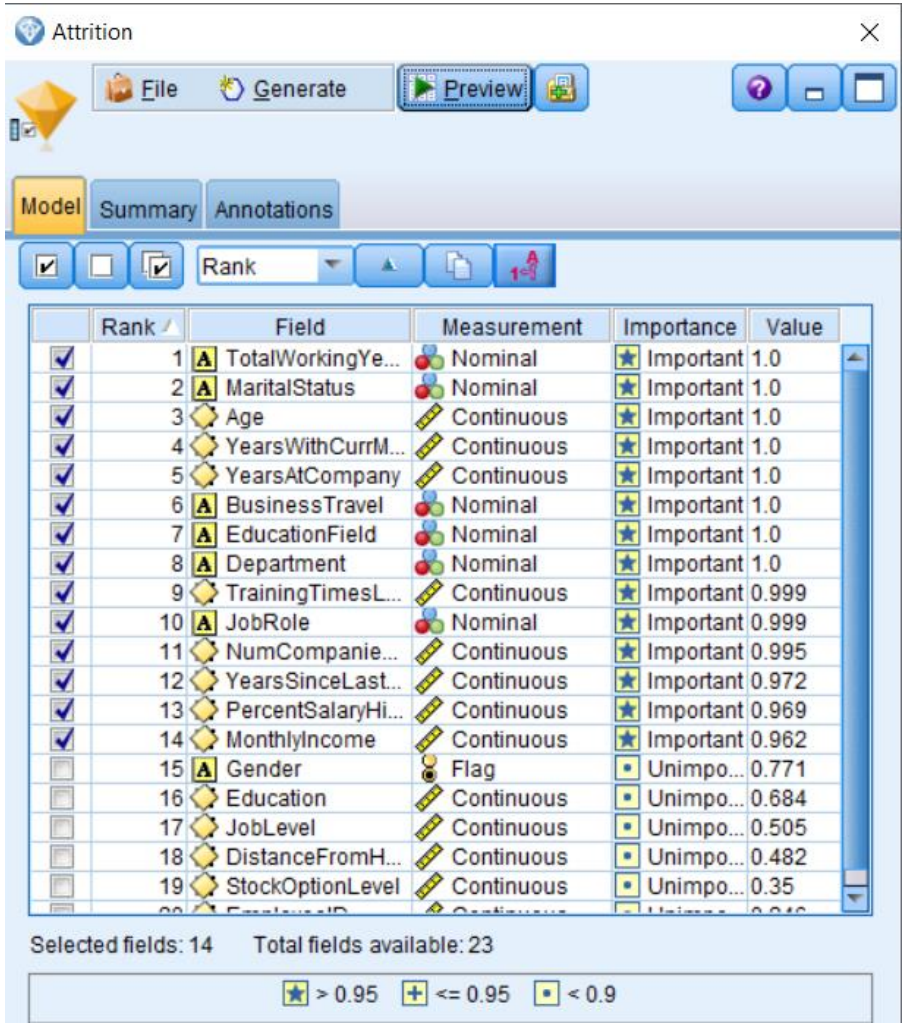
- Data mining problems may involve hundreds, or even thousands, of fields that can potentially be used as inputs. As a result, a great deal of time and effort may be spent examining which fields or variables to include in the model. To narrow down the choices, the Feature Selection algorithm can be used to identify the fields that are most important for a given analysis. For example, if you are trying to predict patient outcomes based on a number of factors, which factors are the most likely to be important?

Bütün hüquqlar qorunur.

Feature Selection Node

- Open ‘general_data.csv’
- Connect **Type node**
- Select ‘Attrition’ as Target
- Connect **Feature Selection node** from Modelling palette
- Run the Stream
- Double click to model nugget





Rank	Field	Measurement	Importance	Value
1	TotalWorkingYe...	Nominal	Important	1.0
2	MaritalStatus	Nominal	Important	1.0
3	Age	Continuous	Important	1.0
4	YearsWithCurrM...	Continuous	Important	1.0
5	YearsAtCompany	Continuous	Important	1.0
6	BusinessTravel	Nominal	Important	1.0
7	EducationField	Nominal	Important	1.0
8	Department	Nominal	Important	1.0
9	TrainingTimesL...	Continuous	Important	0.999
10	JobRole	Nominal	Important	0.999
11	NumCompanie...	Continuous	Important	0.995
12	YearsSinceLast...	Continuous	Important	0.972
13	PercentSalaryHi...	Continuous	Important	0.969
14	MonthlyIncome	Continuous	Important	0.962
15	Gender	Flag	Unimpo...	0.771
16	Education	Continuous	Unimpo...	0.684
17	JobLevel	Continuous	Unimpo...	0.505
18	DistanceFromH...	Continuous	Unimpo...	0.482
19	StockOptionLevel	Continuous	Unimpo...	0.35

Bütün hüquqlar qorunur.

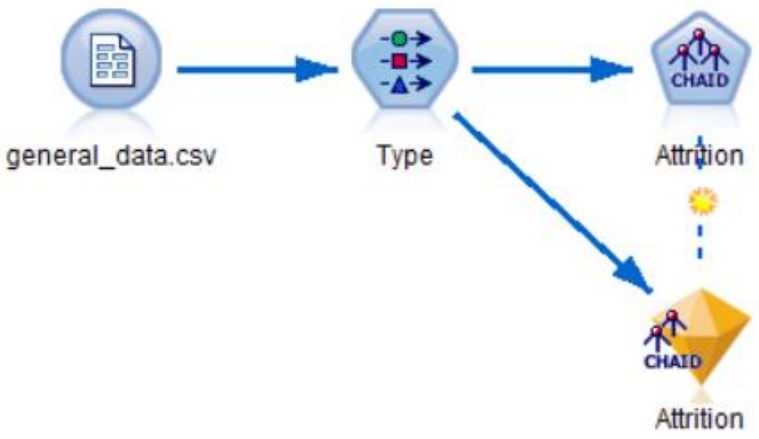
CHAID Node

- CHAID first examines the crosstabulations between each of the input fields and the outcome, and tests for significance using a chi-square independence test. If more than one of these relations is statistically significant, CHAID will select the input field that is the most significant (smallest p value). If an input has more than two categories, these are compared, and categories that show no differences in the outcome are collapsed together. This is done by successively joining the pair of categories showing the least significant difference. This category-merging process stops when all remaining categories differ at the specified testing level. For nominal input fields, any categories can be merged; for an ordinal set, only contiguous categories can be merged.

Bütün hüquqlar qorunur.

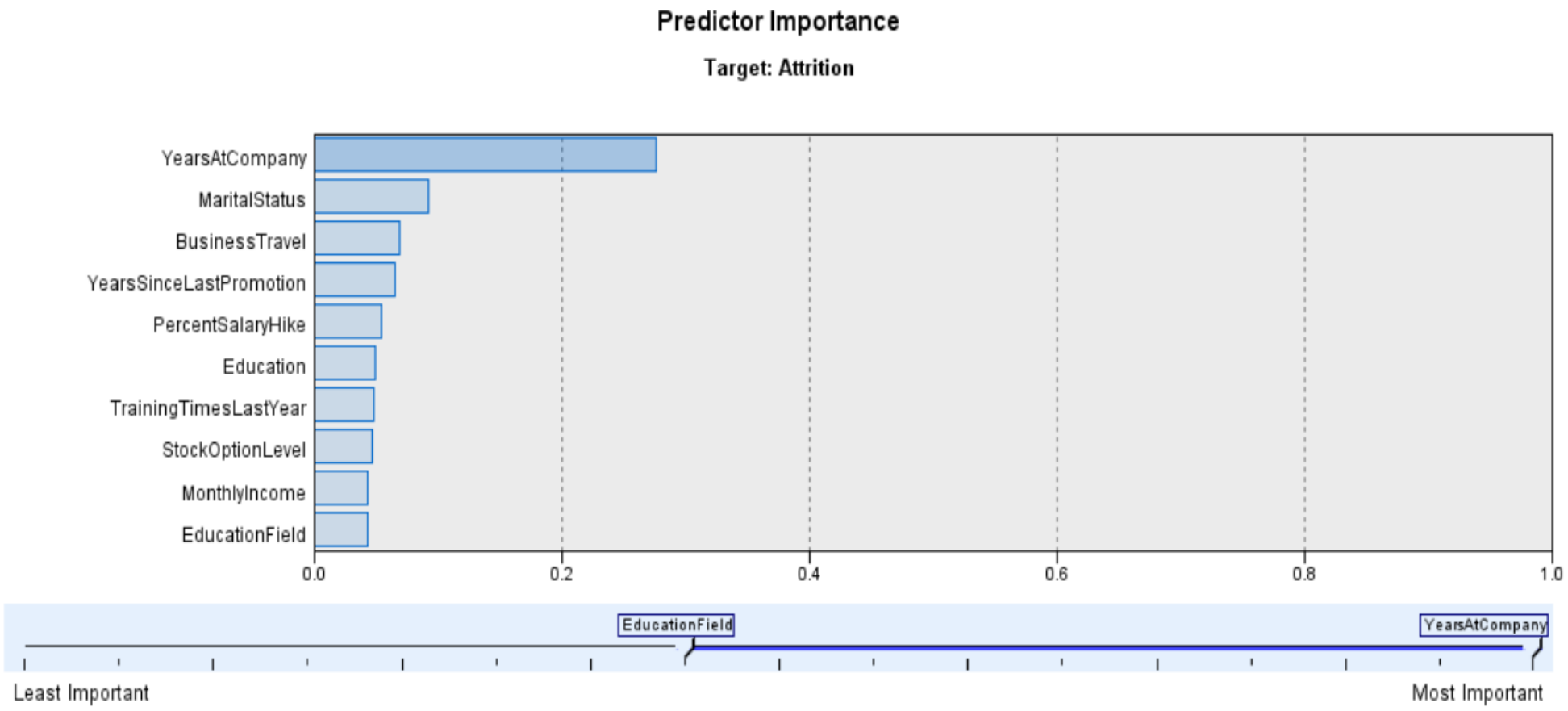
CHAID Node

- Open ‘general_data.csv’
- Connect **Type node**
- Select ‘Attrition’ as Target
- Connect CHAID node from Modelling palette
- Run the stream
- Double click the nugget



Bütün hüquqlar qorunur.

Output

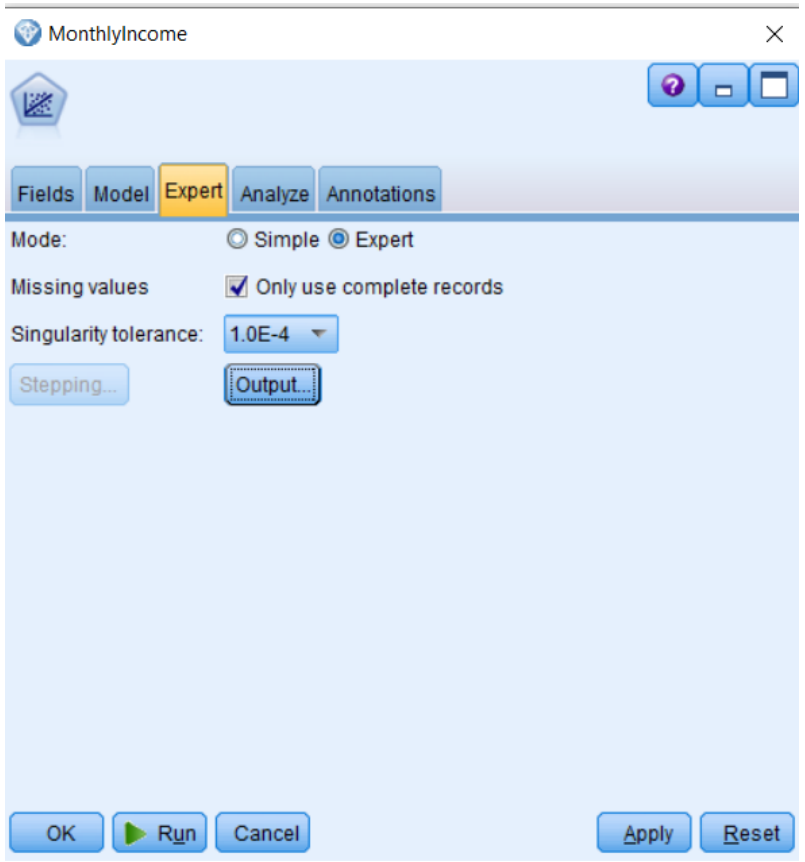


Bütün hüquqlar qorunur.

Regression Node

Regression node provides correlation coefficients and significance levels between numerical variables.

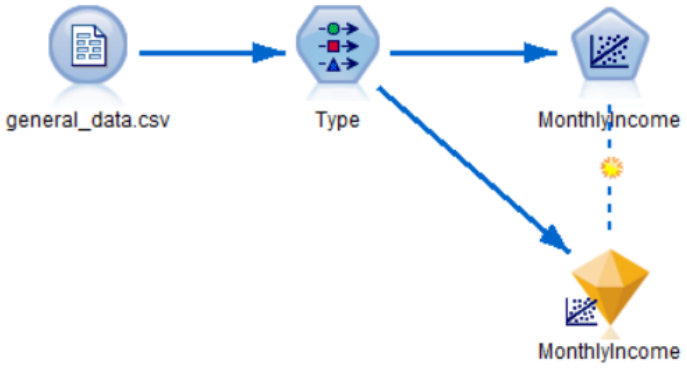
- Open ‘general_data.csv’
- Connect **Type node**
- Select ‘MonthlyIncome’ as Target
- Connect **Regression node** from Modelling palette
- Double click the node
- Come to the **Expert tab**
- Change mode to **Expert**

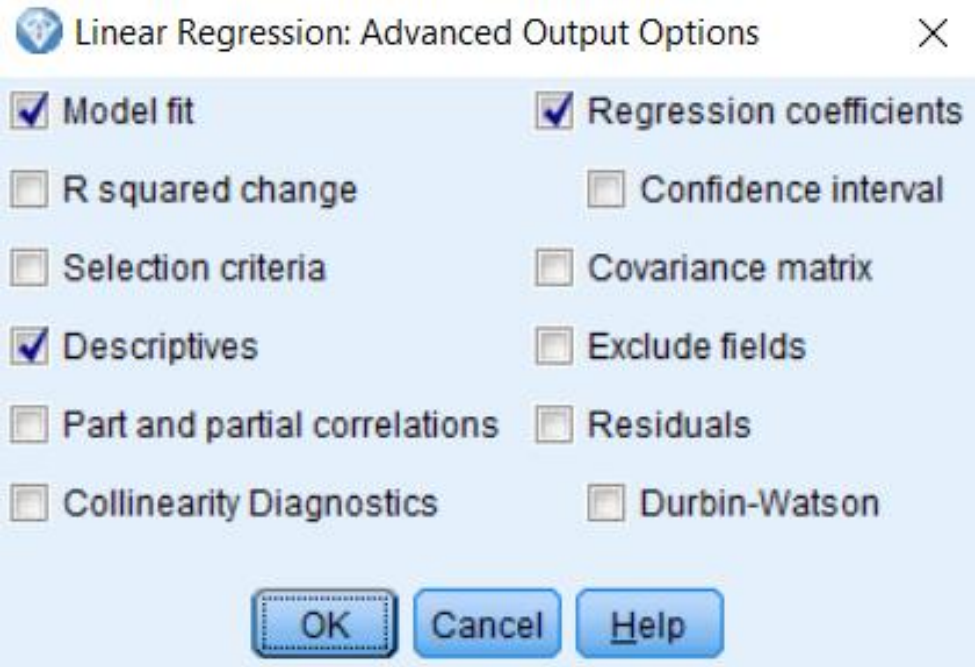


Bütün hüquqlar qorunur.

Regression Node

- Click **Output**
- Select **Descriptives**
- Click OK and Run the stream





Bütün hüquqlar qorunur.

Output

Correlations

		Age	DistanceFromHome	Education	EmployeeCount	EmployeeID	JobLevel	NumCompaniesWorked	PercentSalaryHike	StandardHours	StockOptionLevel	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager	MonthlyIncome
Pearson Correlation	Age	1.000	.008	-.034	.	.007	-.002	.299	-.033	.	-.031	-.029	.311	.216	.202	-.045
	DistanceFromHome	.008	1.000	-.008	.	.000	-.039	-.014	.037	.	.009	-.008	.032	.002	.022	-.022
	Education	-.034	-.008	1.000	.	-.010	.045	-.016	-.041	.	.002	.011	.007	.024	.006	.008
	EmployeeCount	.	.	.	1.000	.	.	.	.	.	.	.	.	.	.	.
	EmployeeID	.007	.000	-.010	.	1.000	-.002	-.001	-.005	.	-.014	-.011	.004	.000	.009	.007
	JobLevel	-.002	-.039	.045	.	-.002	1.000	-.010	.011	.	.001	-.033	-.064	-.060	-.054	.047
	NumCompaniesWorked	.299	-.014	-.016	.	-.001	-.010	1.000	.031	.	.017	-.032	-.118	-.037	-.110	-.021
	PercentSalaryHike	-.033	.037	-.041	.	-.005	.011	.031	1.000	.	.012	-.037	-.029	-.029	-.040	.005
	StandardHours	.	.	.	.	.	.	.	.	1.000	.	.	.	.	.	.
	StockOptionLevel	-.031	.009	.002	.	-.014	.001	.017	.012	.	1.000	-.069	.008	.019	.018	.027
	TrainingTimesLastYear	-.029	-.008	.011	.	-.011	-.033	-.032	-.037	.	-.069	1.000	-.008	.016	-.014	.050
	YearsAtCompany	.311	.032	.007	.	.004	-.064	-.118	-.029	.	.008	-.008	1.000	.619	.769	.001
	YearsSinceLastPromotion	.216	.002	.024	.	.000	-.060	-.037	-.029	.	.019	.016	.619	1.000	.510	.065
	YearsWithCurrManager	.202	.022	.006	.	.009	-.054	-.110	-.040	.	.018	-.014	.769	.510	1.000	.024
	MonthlyIncome	-.045	-.022	.008	.	.007	.047	-.021	.005	.	.027	.050	.001	.065	.024	1.000
Sig. (1-tailed)	Age	.	.308	.013	.000	.319	.456	.000	.014	.000	.018	.029	.000	.000	.000	.001
	DistanceFromHome	.308	.	.298	.000	.496	.005	.180	.007	.000	.265	.292	.018	.443	.072	.073
	Education	.013	.298	.	.000	.253	.001	.141	.003	.000	.456	.235	.329	.057	.335	.305
	EmployeeCount	.000	.000	.000	.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	EmployeeID	.319	.496	.253	.000	.	.441	.471	.381	.000	.178	.224	.388	.499	.272	.322
	JobLevel	.456	.005	.001	.000	.441	.	.256	.239	.000	.482	.015	.000	.000	.000	.001
	NumCompaniesWorked	.000	.180	.141	.000	.471	.256	.	.020	.000	.128	.016	.000	.008	.000	.085
	PercentSalaryHike	.014	.007	.003	.000	.381	.239	.020	.	.000	.215	.007	.026	.025	.004	.380
	StandardHours	.000	.000	.000	.000	.000	.000	.000	.000	.	.000	.000	.000	.000	.000	.000
	StockOptionLevel	.018	.265	.456	.000	.178	.482	.128	.215	.000	.	.000	.299	.101	.114	.038
	TrainingTimesLastYear	.029	.292	.235	.000	.224	.015	.016	.007	.000	.000	.	.292	.146	.178	.000
	YearsAtCompany	.000	.018	.329	.000	.388	.000	.000	.026	.000	.299	.292	.	.000	.000	.487
	YearsSinceLastPromotion	.000	.443	.057	.000	.499	.000	.008	.025	.000	.101	.146	.000	.	.000	.000
	YearsWithCurrManager	.000	.072	.335	.000	.272	.000	.000	.004	.000	.114	.178	.000	.000	.	.059
	MonthlyIncome	.001	.073	.305	.000	.322	.001	.085	.380	.000	.038	.000	.487	.000	.059	.

Bütün hüquqlar qorunur.

II
Machine Learning using SPSS Modeler

Data Science Academy

Content

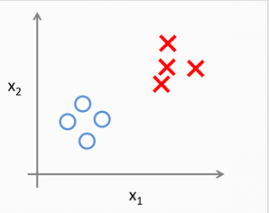
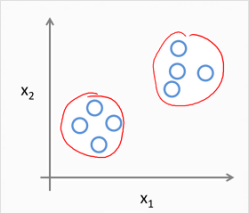
- Supervised Learning
  - Classification Models
    - CHAID modelling
    - SVM algorithm
    - CART modeling
    - Logistic Regression
    - Auto Classifier
  - Regression Models
    - GLM
    - Auto Numeric
    - Linear Regression
    - LSVM
    - Random Trees
    - Generalized Linear
  - Ensemble Modeling

- Unsupervised Learning
  - Cluster Analysis
    - K- Means Clustering
    - Two – Step Clustering
    - Kahonen Network
    - Anomaly algorithm – Fraud Detection
    - Auto Clustering
  - Association Rule
    - Apriori algorithm – Market Basket Analysis
    - Sequence algorithm

Bütün hüquqlar qorunur.

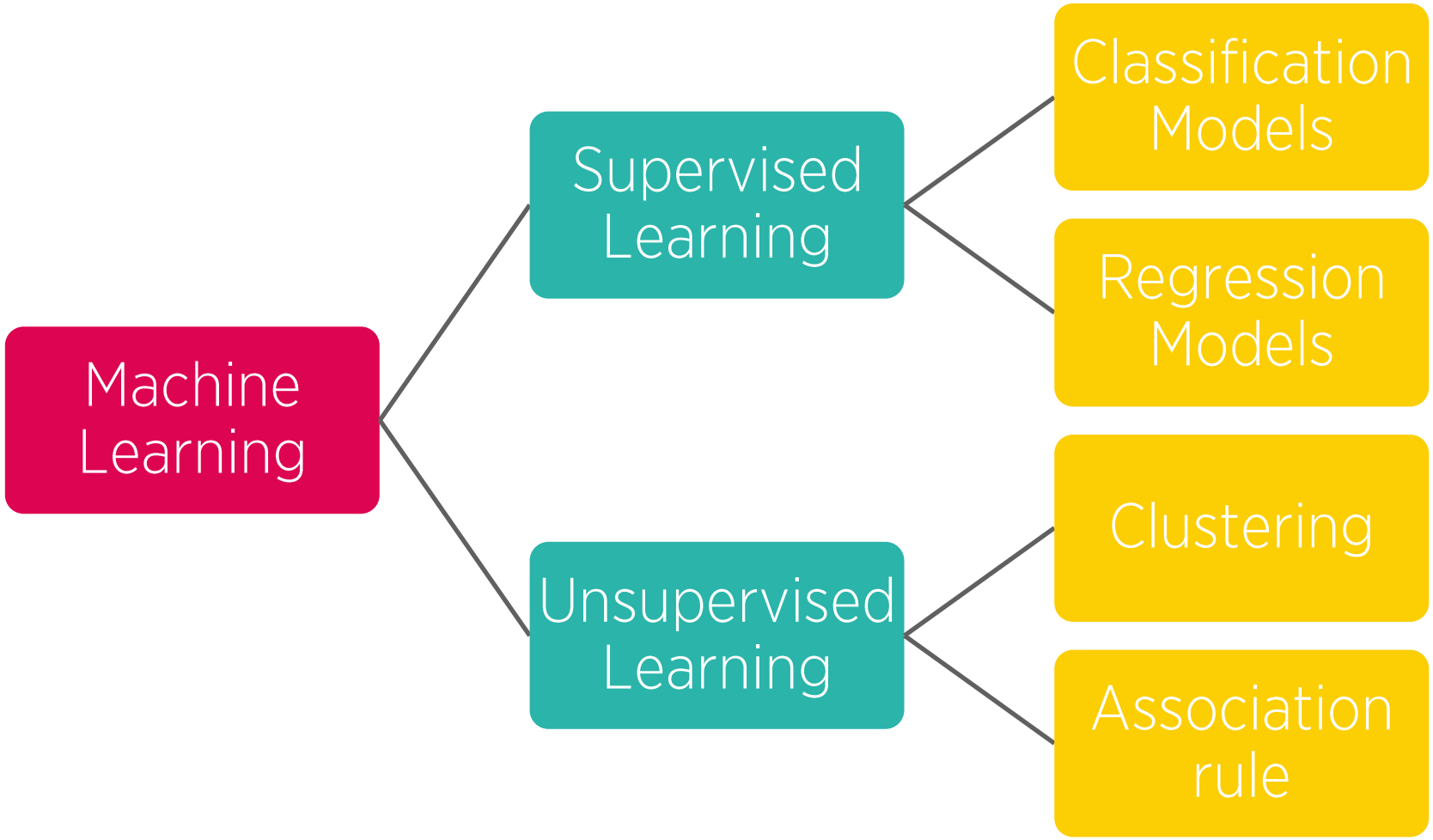
Supervised & Unsupervised Learning

- Supervised Learning provides predictive modeling using dependent and independent variables. The regression and classification models belong to this group.
- Unsupervised Learning provides the interpretation of data by creating different groups using independent variables. K-Means Clustering, Two-Step, Anomaly Detection and such algorithms refer to this group.

Supervised	Unsupervised
Input Data is labelled	Input Data is unlabeled
Uses training dataset	Uses just input dataset
Data is classified based on training dataset	Uses properties of given data to classify it
Used for prediction	Used for analysis
Divided into two types: Regression & Classification	Divided into two types: Clustering & Association
Known number of classes	Unknown number of classes
	
Use off-line analysis of data	Use real-time analysis of data

Bütün hüquqlar qorunur.

Supervised Learning & Unsupervised Learning



Bütün hüquqlar qorunur.

SUPERVISED LEARNING

Data Science Academy

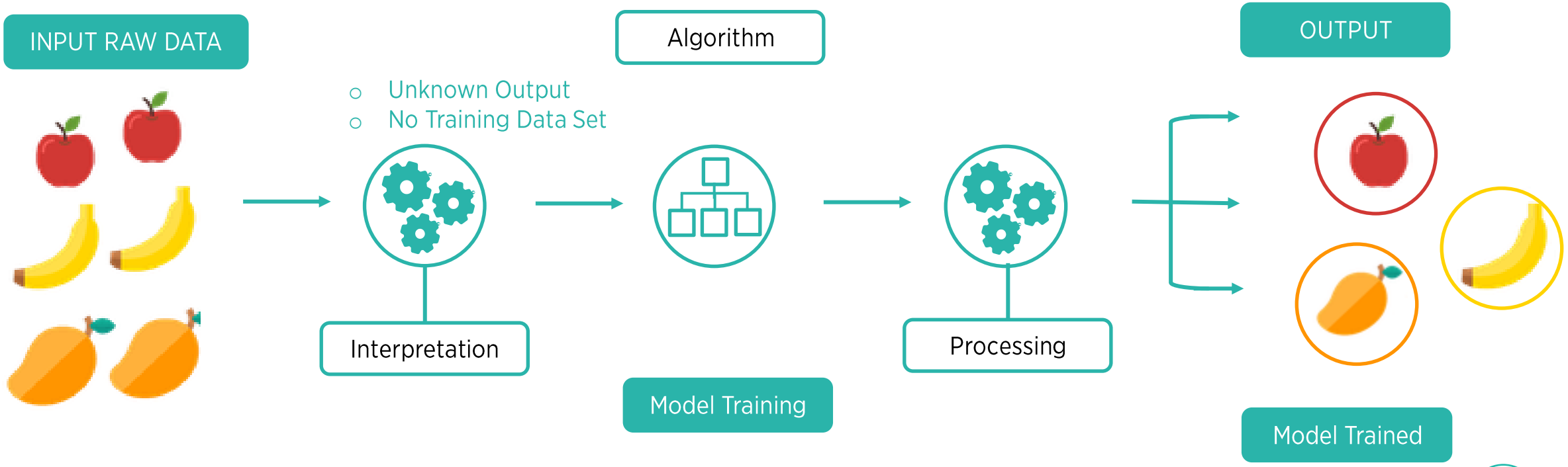
Supervised Learning

- A **supervised machine learning** algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.
- Imagine a computer is a child, we are its supervisor (e.g. parent, guardian, or teacher), and we want the child (computer) to learn what a pig looks like. We will show the child several different pictures, some of which are pigs and the rest could be pictures of anything (cats, dogs, etc).
- When we see a pig, we shout "pig!" When it's not a pig, we shout "no, not pig!" After doing this several times with the child, we show them a picture and ask "pig?" and they will correctly (most of the time) say "pig!" or "no, not pig!" depending on what the picture is. That is supervised machine learning.
- Supervised machine learning algorithms are used to solve **classification** or **regression** problems.

Bütün hüquqlar qorunur.

Classification Model

- A **classification** problem has a discrete value as its output. For example, “likes pineapple on pizza” and “does not like pineapple on pizza” are discrete. There is no middle ground. The analogy above of teaching a child to identify a pig is another example of a classification problem.



Bütün hüquqlar qorunur.

Classification Model

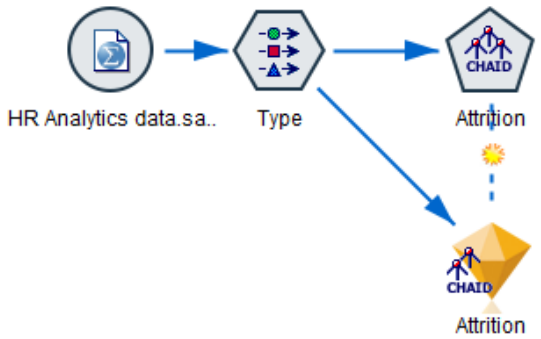
- This image shows a basic example of what classification data might look like. We have a predictor (or set of predictors) and a label. In the image, we might be trying to predict whether someone likes pineapple (1) on their pizza or not (0) based on their age (the predictor).

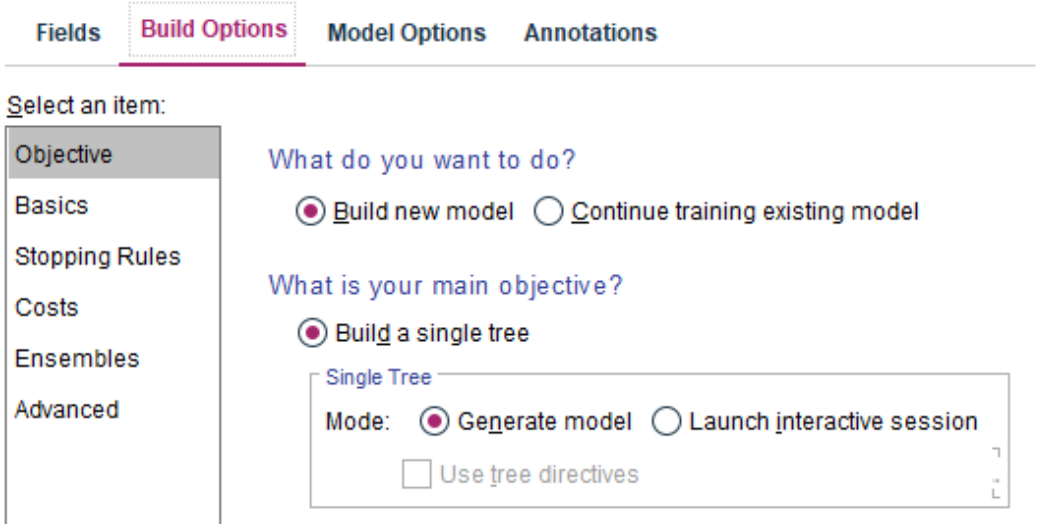
Age	Likes Pineapple on Pizza
42	1
65	1
50	1
76	1
96	1
50	1
91	0
58	1
25	1
23	1
75	1

Bütün hüquqlar qorunur.

CHAID Modelling

1. Open “HR Analytics data.sav”;
2. Connect **Type** node;
3. **Role:** Atrition > Target;
4. Connect CHAID node from Modelling palette:
   - Open **Build options** tab
   - Select **Objective**
   - **Mode**: Generate Model

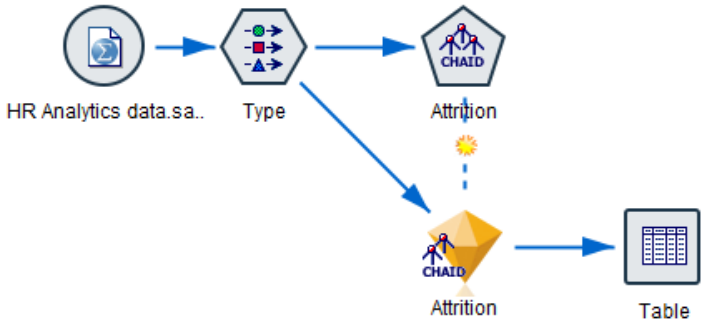


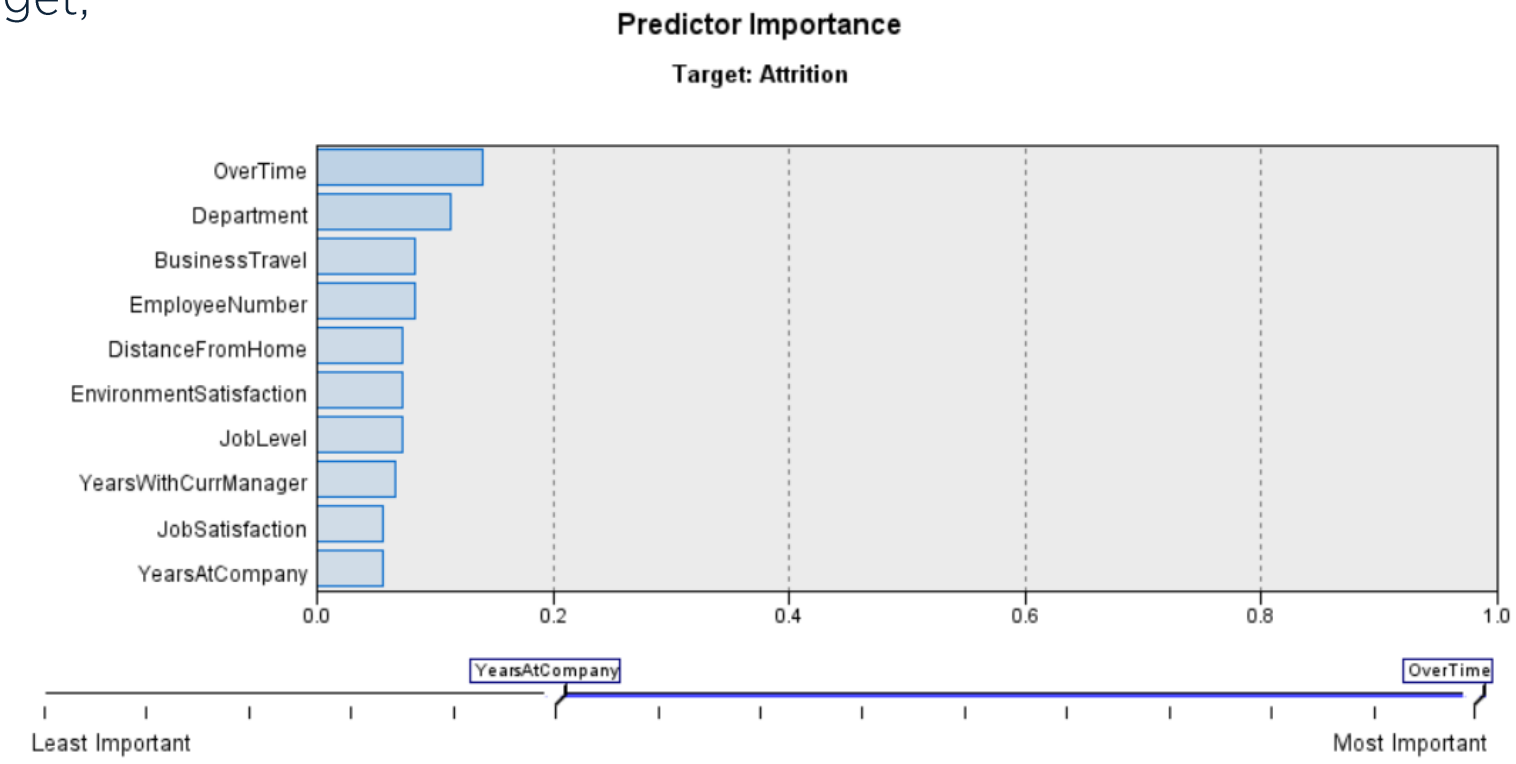


Bütün hüquqlar qorunur.

Output

1. Connect **Table node** to the Nugget;





Bütün hüquqlar qorunur.

Output

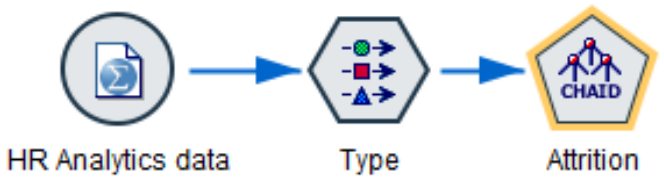
Table Annotations

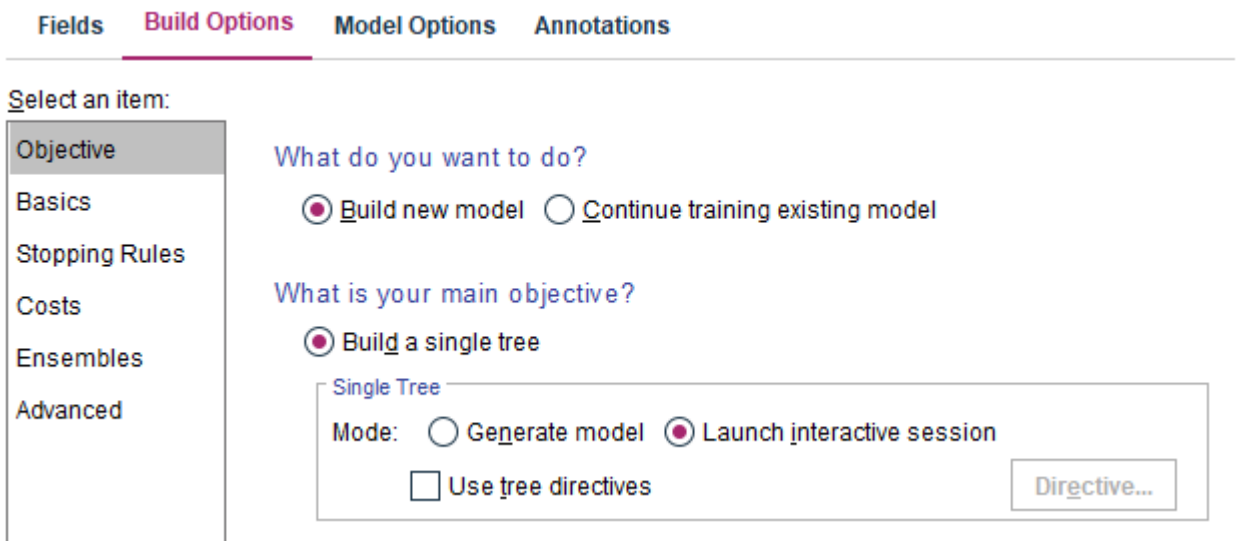
	Age	Attrition	BusinessTravel	DailyRate	Department	$R-Attrition	$RC-Attrition
1	20.000	Yes	Travel_Rarely	1362.000	Resea	Yes	0.91
2	18.000	No	Non-Travel	287.000	Resea	No	0.67
3	28.000	No	Travel_Rarely	1144.000	Sales	No	0.67
4	30.000	Yes	Travel_Rarely	945.000	Sales	No	0.67
5	29.000	Yes	Travel_Freque	746.000	Sales	Yes	0.70
6	19.000	Yes	Travel_Rarely	303.000	Resea	No	0.67
7	25.000	Yes	Travel_Freque	599.000	Sales	Yes	0.91
8	31.000	No	Travel_Rarely	1276.000	Resea	No	0.81
9	18.000	No	Travel_Rarely	812.000	Sales	No	0.67
10	23.000	No	Travel_Rarely	373.000	Resea	No	0.95
11	21.000	No	Travel_Rarely	391.000	Resea	No	0.67
12	31.000	Yes	Travel_Rarely	202.000	Resea	No	0.67
13	30.000	No	Non-Travel	990.000	Resea	No	0.67
14	35.000	No	Non-Travel	727.000	Resea	No	0.98
15	31.000	Yes	Travel_Freque	667.000	Sales	Yes	0.70
16	32.000	Yes	Travel_Rarely	1259.000	Resea	No	0.78
17	21.000	Yes	Travel_Rarely	1334.000	Resea	No	0.67
18	18.000	Yes	Travel_Rarely	230.000	Resea	No	0.67
19	19.000	No	Travel_Rarely	1181.000	Resea	No	0.67
20	18.000	No	Non-Travel	1431.000	Resea	No	0.67
21	24.000	Yes	Travel_Rarely	240.000	Human	No	0.67

Bütün hüquqlar qorunur.

CHAID Modelling

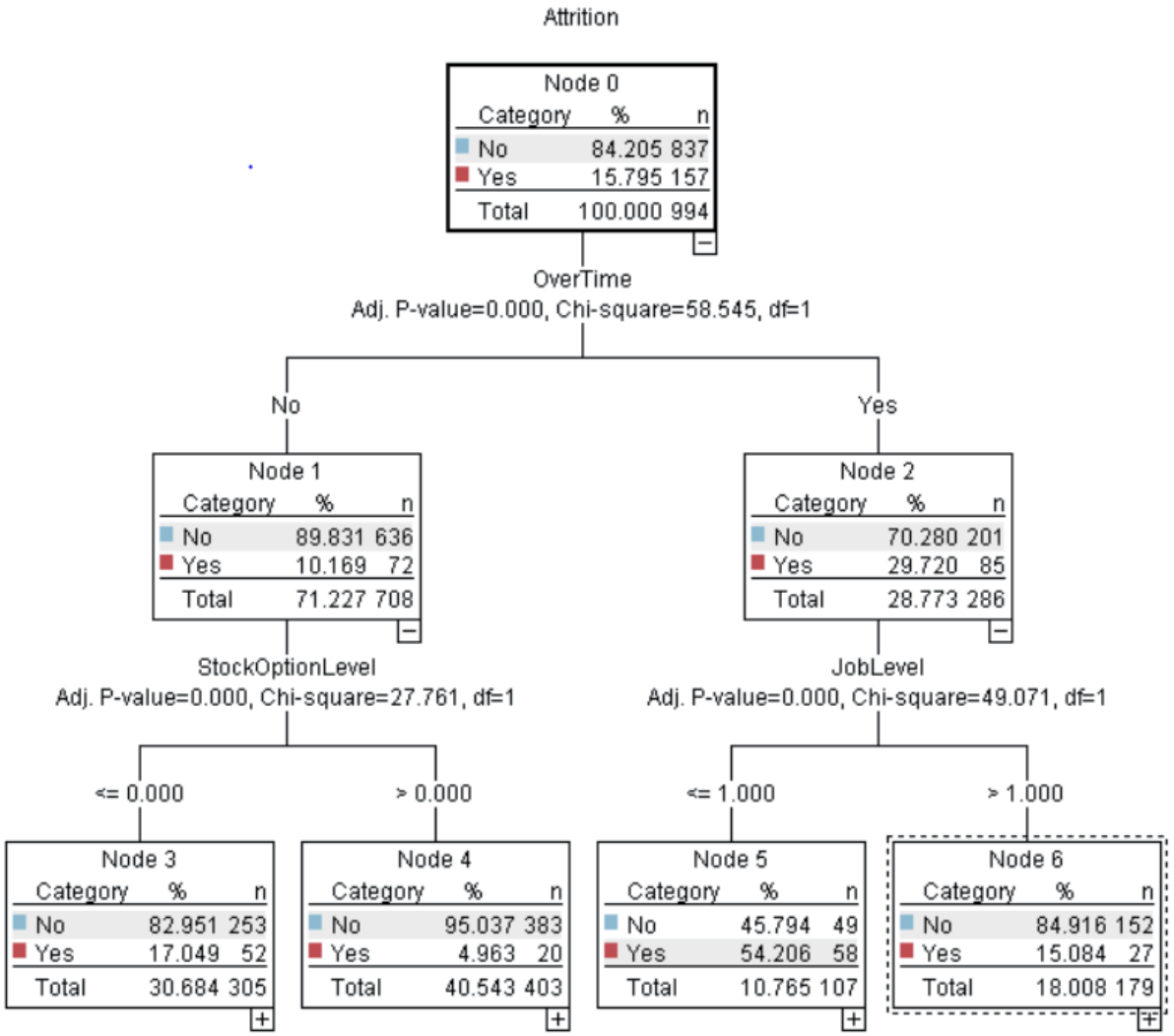
1. Open “HR Analytics data.sav”;
2. Connect **Type** node;
3. **Role:** Atrition > Target;
4. Connect CHAID node from Modelling palette:
   - Open **Build options** tab
   - Select **Objective**
   - **Mode**: Launch interactive session





Bütün hüquqlar qorunur.

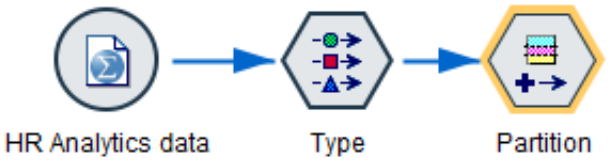
Output

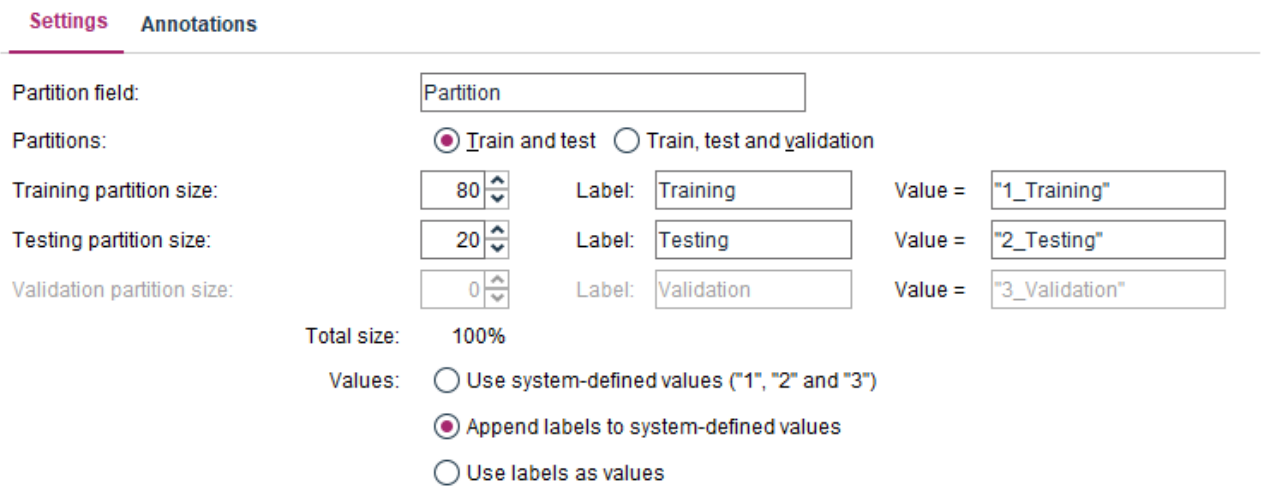


Bütün hüquqlar qorunur.

CHAID Modelling with Partition

1. Open *"HR Analytics data.sav"*;
2. Connect **Type** node
3. Target: Atrition
4. Add **Partition** node from **Field Ops** palette:
   Training partition size: 80
   Testing partition size: 20
5. Preview

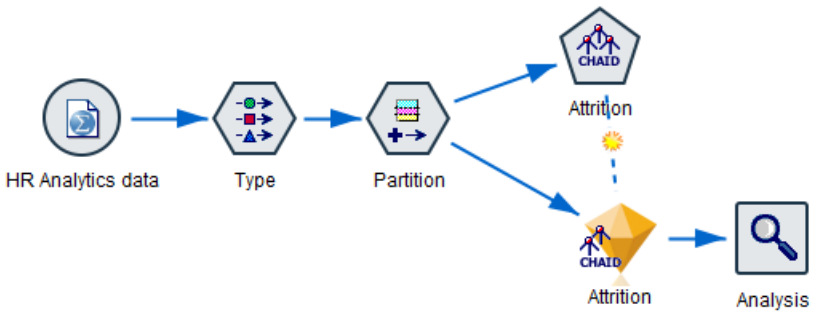




Bütün hüquqlar qorunur.

CHAID Modelling Results

1. CHAID> Build Options> Generate model
2. Connect **Analysis** node from **Output** palette to **Nugget**
3. Run



Analysis Annotations

Collapse All Expand All

Results for output field Attrition

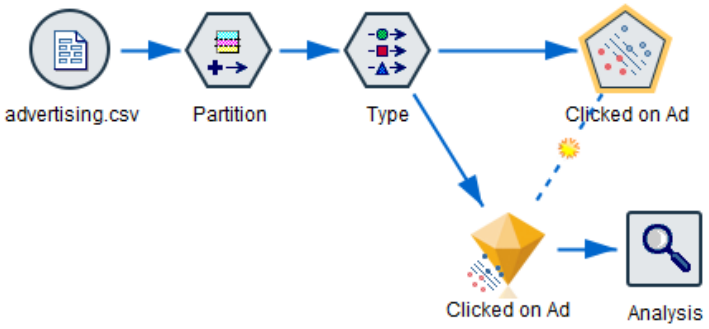
Comparing $R-Attrition with Attrition

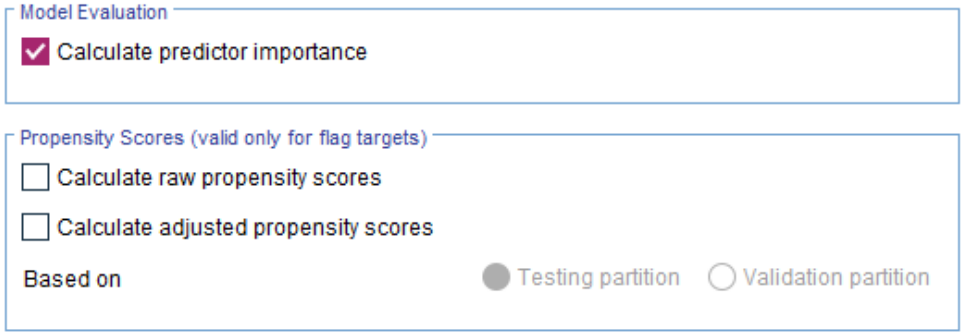
'Partition'	1_Training		2_Testing	
Correct	1,021	87.19%	257	85.95%
Wrong	150	12.81%	42	14.05%
Total	1,171		299	

Bütün hüquqlar qorunur.

Support Vector Machine(SVM) algorithm

1. Open *"advertising.csv"*;
2. Connect **Partition node** and set partition sizes - Training:70, Testing:30
3. Add **Type** node and set **Role:**
   Clicked on Ad (Measurement – Flag) -> **Target**;
   Daily Time Spent; Age; Area Income; Daily internet usage; Male(Measurement-flag) -> **Input**.
4. Add **SVM** node from **Modelling** palette
   - **Analyse**: Enable **Calculate predictor importance**
5. Run and **Analysis node** from **Output palette**. Enable **Coincidence matrices** from **Edit menu** and run.





Bütün hüquqlar qorunur.

Output

Analysis Annotations

Collapse All Expand All

Results for output field Clicked on Ad

Comparing $S-Clicked on Ad with Clicked on Ad

'Partition'	1_Training		2_Testing	
Correct	678	96.86%	292	97.33%
Wrong	22	3.14%	8	2.67%
Total	700		300	

Coincidence Matrix for $S-Clicked on Ad (rows show actuals)

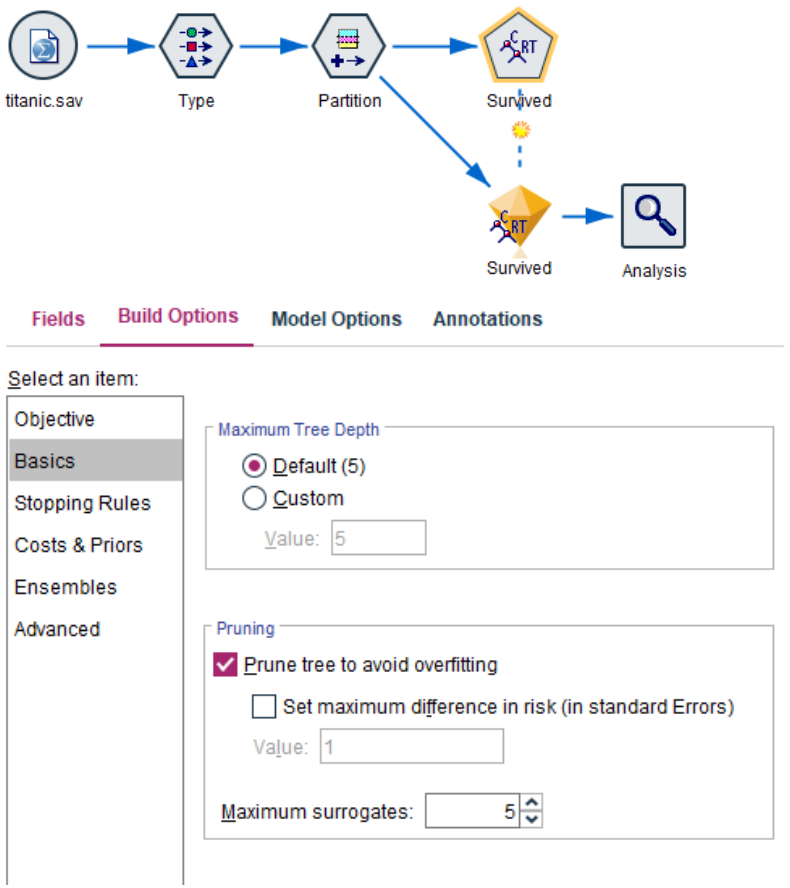
'Partition' = 1_Training	0	1
0	341	7
1	15	337

'Partition' = 2_Testing	0	1
0	150	2
1	6	142

Bütün hüquqlar qorunur.

Classification & Regression Trees (CART) Modelling

1. Open *"titanic.sav"*;
2. Add **Type** node and set **Role**:
   Survived (as Flag) -> **Target**;
   PassengerId, Name, Ticket, Cabin -> **None**;
3. Set **measurement** of Pclass as **ordinal**;
4. Add **Partition** node and set partition sizes - **Training:70**, **Testing:30**;
5. Add **C&RT node** from **Modelling** palette;
6. Run and add **Analysis node** from **Output** palette. Enable **Coincidence matrices** from **Edit menu** and run again.



Bütün hüquqlar qorunur.

Output

Results for output field Survived

Comparing $R-Survived with Survived

'Partition'	1_Training		2_Testing	
Correct	530	84.39%	210	79.85%
Wrong	98	15.61%	53	20.15%
Total	628		263	

Coincidence Matrix for $R-Survived (rows show actuals)

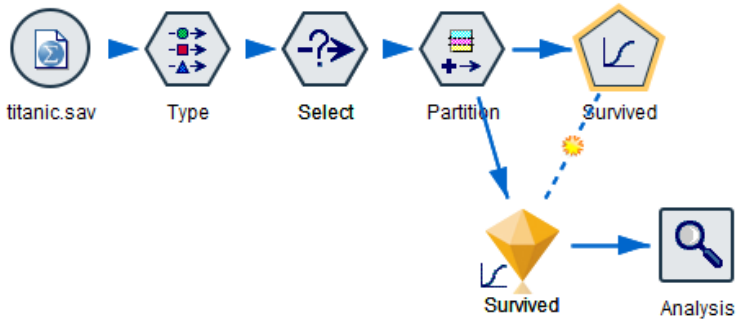
'Partition' = 1_Training	0.000000	1.000000
0.000000	352	37
1.000000	61	178

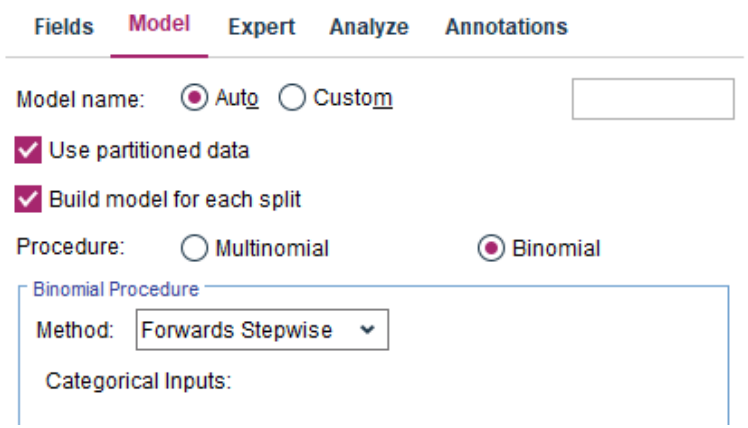
'Partition' = 2_Testing	0.000000	1.000000
0.000000	138	22
1.000000	31	72

Bütün hüquqlar qorunur.

Logistic Regression – Case Titanic

1. Open *“titanic.sav”*;
2. Add **Type** node and set **Role**:
   Survived (as Flag) -> **Target**;
   PassengerId, Name, Ticket, Cabin -> **None**;
3. In **Select node** type: **Age>0**;
4. Add **Partition node** and set partition sizes - **Training:70, Testing:30**;
5. Add **Logistic Regression** node from **Modelling** palette
   - **Procedure**: Binomial
   - **Method:** Forwards Stepwise
6. Run and add **Analysis node** from **Output palette**.
   Enable **Coincidence matrices** from **Edit menu** and run.





Bütün hüquqlar qorunur.

Output

Results for output field Survived

Comparing $L-Survived with Survived

'Partition'	1_Training		2_Testing	
Correct	403	64.17%	162	61.6%
Wrong	225	35.83%	101	38.4%
Total	628		263	

Coincidence Matrix for $L-Survived (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	$null$
0.000000	258	46	85
1.000000	56	145	38

'Partition' = 2_Testing	0.000000	1.000000	$null$
0.000000	101	19	40
1.000000	28	61	14

Bütün hüquqlar qorunur.

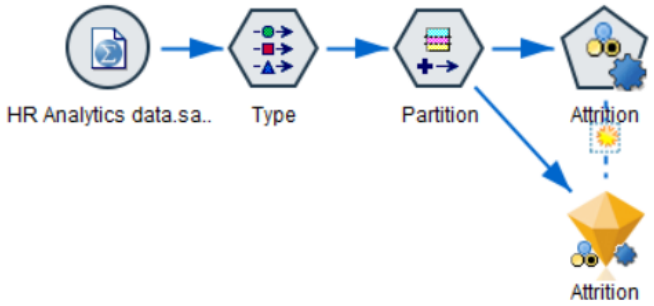
Auto Classifier node

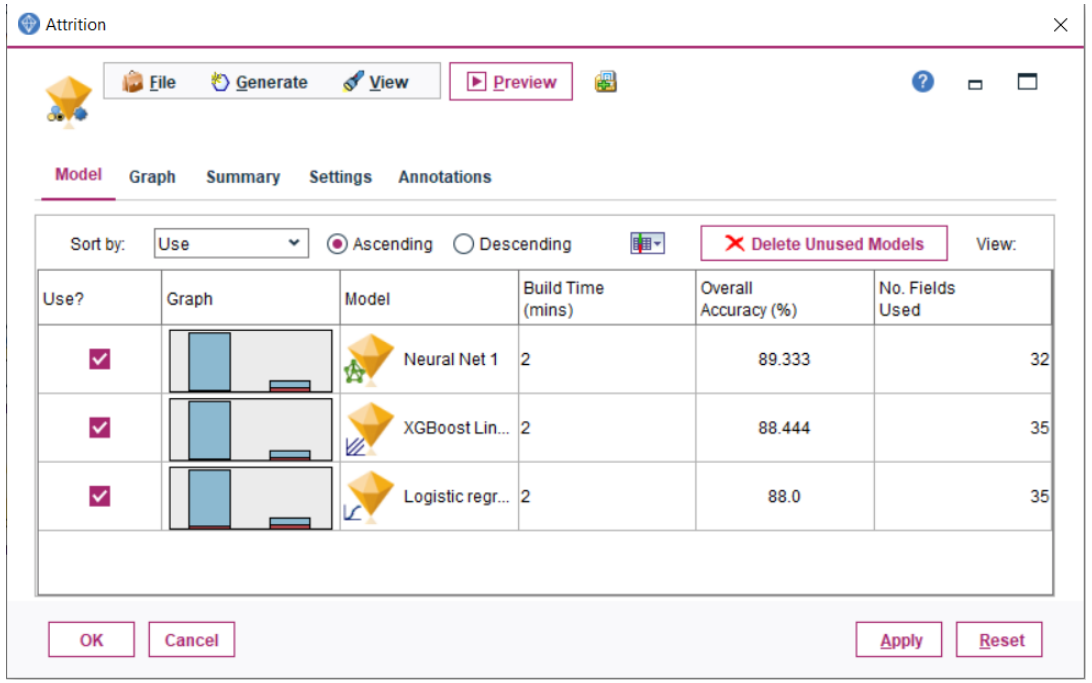
- The Auto Classifier node creates and compares a number of different models for binary outcomes (yes or no, churn or do not churn, and so on), allowing you to choose the best approach for a given analysis. A number of modeling algorithms are supported, making it possible to select the methods you want to use, the specific options for each, and the criteria for comparing the results. The node generates a set of models based on the specified options and ranks the best candidates according to the criteria you specify.

Bütün hüquqlar qorunur.

Auto Classifier

1. Open *"HR analytics.sav"*;
2. Connect **Type** node
   Role: Attrition -> Target;
3. Add Partition node;
4. Connect **Auto Classifier** node from **Modelling** palette
5. Run





Bütün hüquqlar qorunur.

III

Model Evaluation and Analysis using SPSS Modeler

Data Science Academy

Evaluation node

- The Evaluation node offers an easy way to evaluate and compare predictive models to choose the best model for your application. Evaluation charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (quantiles), and then plotting the value of the business criterion for each quantile, from highest to lowest. Multiple models are shown as separate lines in the plot.

Bütün hüquqlar qorunur.

Classification Model Evaluation methods

- Confusion matrix
- Precision
- Recall
- F-1 score
- ROC
- AUC
- Gini Coefficient

Bütün hüquqlar qorunur.

Confusion Matrix Parameters

- True Positives (TP) – These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.
- True Negatives (TN) – These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.
- False Positives (FP) –When actual class is no and predicted class is yes, E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.
- False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Bütün hüquqlar qorunur.

Confusion Matrix Parameters

Predicted Class

Actual Class

	Class = Yes	Class = No
Class = Yes	True positive	False negative
Class = No	False positive	True negative

Bütün hüquqlar qorunur.

Performance Measures

- **Precision** – Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

  Precision = TP/(TP+FP)

- **Recall (Sensitivity)** – Recall is the ratio of correctly predicted positive observations to the all observations in actual class – yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for his model as it's above 0.5.

  Recall = TP/(TP+FN)

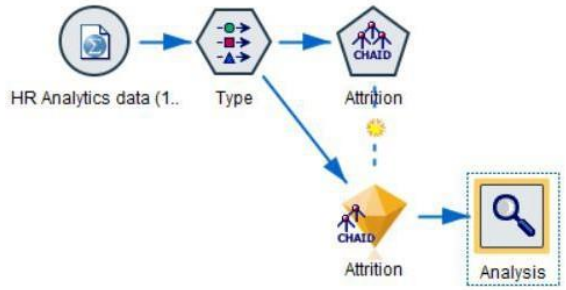
- **F1 score** – F1 – F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account . Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially of you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look both Precision and Recall. In our case, F1 score is 0.701.

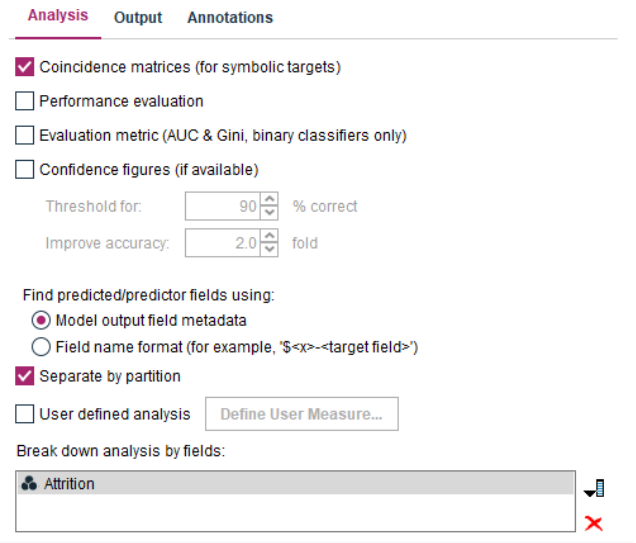
  F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Bütün hüquqlar qorunur.

Recall

- Build **CHAID model** on *"HR Analytics.sav"* dataset as shown in stream
- Add **Analysis node** from **Output palette**
- Choose *Attrition* as **Break down analysis by fields**
- Run the stream





Results for output field Attrition

Overall Results

Comparing $R-Attrition with Attrition

Correct	1,284	87.35%
Wrong	186	12.65%
Total	1,470	

Coincidence Matrix for $R-Attrition (rows show actuals)

	No	**Yes**
No	1,202	31
Yes	155	82

Output field Attrition, splitting by field Attrition

Attrition = No

Comparing $R-Attrition with Attrition

Correct	1,202	97.49%
Wrong	31	2.51%
Total	1,233	

Coincidence Matrix for $R-Attrition (rows show actuals)

Attrition = Yes

Comparing $R-Attrition with Attrition

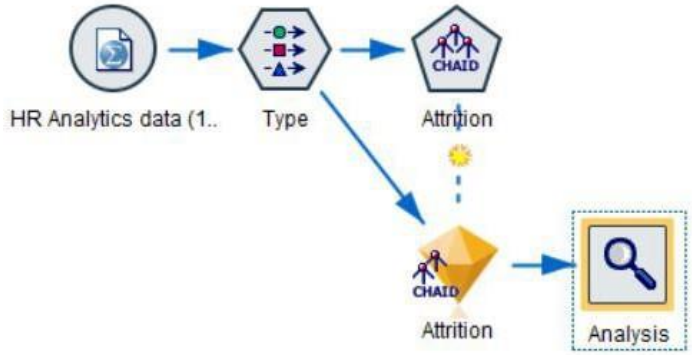
Correct	82	34.6%
Wrong	155	65.4%
Total	237	

Coincidence Matrix for $R-Attrition (rows show actuals)

Bütün hüquqlar qorunur.

Precision

- Build **CHAID model** on *"HR Analytics.sav"* dataset as shown in stream
- Add **Analysis node** from **Output palette**
- Choose *$R-Attrition* as **Break down analysis by fields**
- Run the stream



Results for output field Attrition

Overall Results

Comparing $R-Attrition with Attrition

Correct	1,284	87.35%
Wrong	186	12.65%
Total	1,470	

Coincidence Matrix for $R-Attrition (rows show actuals)

	No	Yes
No	1,202	31
Yes	155	82

Output field Attrition, splitting by field $R-Attrition

$R-Attrition = No

Comparing $R-Attrition with Attrition

Correct	1,202	88.58%
Wrong	155	11.42%
Total	1,357	

Coincidence Matrix for $R-Attrition (rows show actuals)

$R-Attrition = Yes

Comparing $R-Attrition with Attrition

Correct	82	72.57%
Wrong	31	27.43%
Total	113	

Coincidence Matrix for $R-Attrition (rows show actuals)

Bütün hüquqlar qorunur.

ROC and AUC

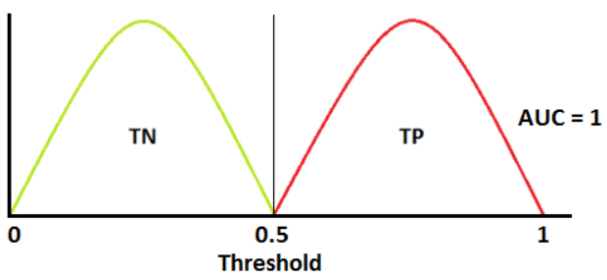
- AUC – ROC curve is a performance measurement for classification problem at various thresholds setting. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguish between classes. Higher the AUC better the model is at distinguishing between patients with disease and no disease.

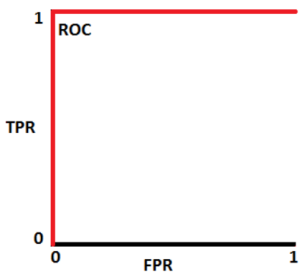
- An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

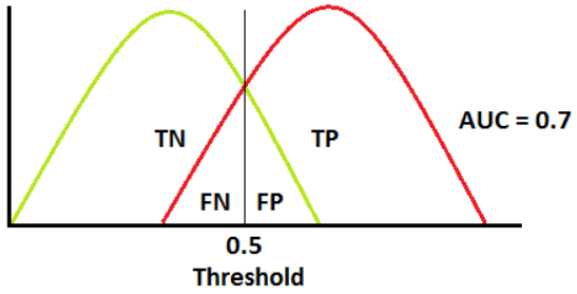
Bütün hüquqlar qorunur.

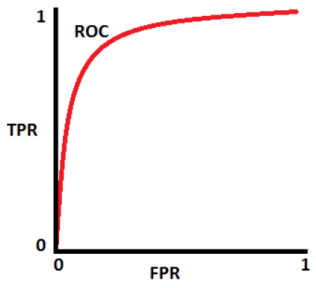
ROC and AUC

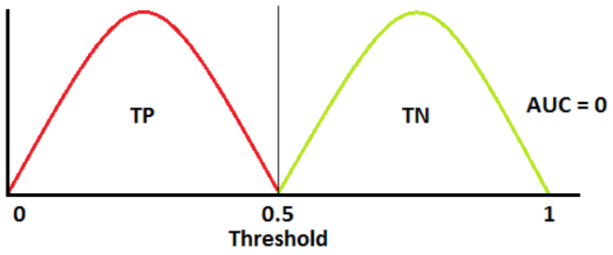
- Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class (patients with no disease)

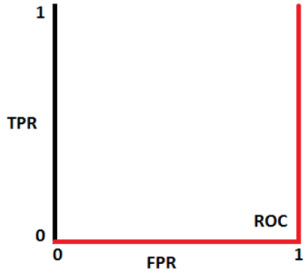


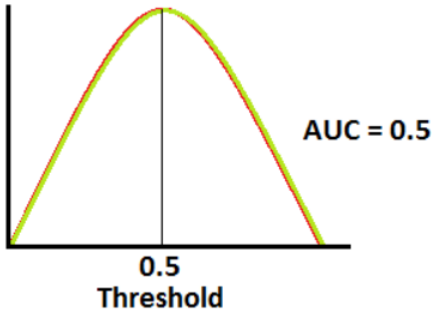


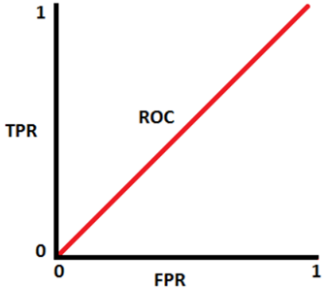












Bütün hüquqlar qorunur.

Analysis node

- The Analysis node allows you to evaluate the ability of a model to generate accurate predictions. Analysis nodes perform various comparisons between predicted values and actual values (your target field) for one or more model nuggets. Analysis nodes can be also used to compare predictive models to other predictive models.

- When you execute an Analysis node, a summary of the analysis results is automatically added to the Analysis section on the Summary tab for each model nugget in the executed stream. The detailed analysis results appear on the Outputs tab of the manager window or can be written directly to a file.

Bütün hüquqlar qorunur.

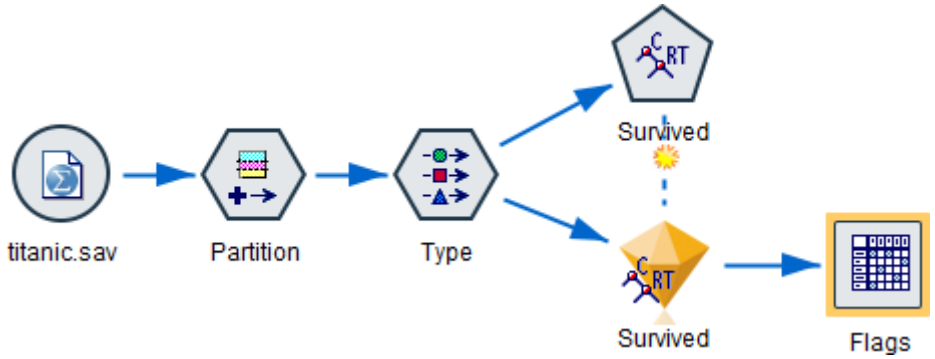
Analysis node

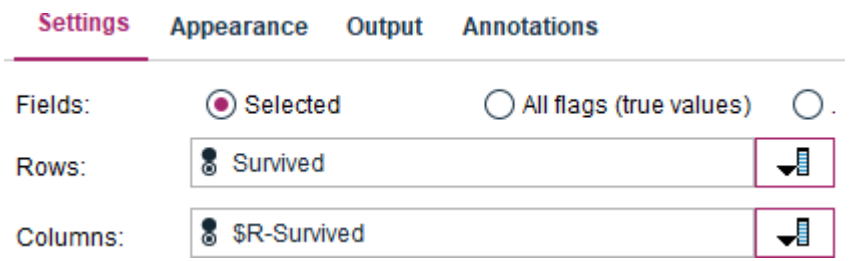
- **Coincidence matrices (for symbolic or categorical targets)** show the pattern of matches between each generated (predicted) field and its target field for categorical targets (either flag, nominal, or ordinal).
- **The AUC evaluation metric** is calculated as the area under an ROC (receiver operator characteristic) curve, and is a scalar representation of the expected performance of a classifier. The AUC is always between 0 and 1, with a higher number representing a better classifier.
- **The Gini coefficient** evaluation metric is sometimes used as an alternative evaluation metric to the AUC, and the two measures are closely related. The Gini coefficient is calculated as twice the area between the ROC curve and the diagonal, or as Gini = 2AUC - 1.
- **Confidence figures (if available)** for models that generate a confidence field, this option reports statistics on the confidence values and their relationship to predictions. There are two settings for this option.

Bütün hüquqlar qorunur.

Confusion matrix

1. Open titanic partitioned.str stream file
2. Add **C&RT node** from Modeling palette
3. Run stream
4. Add matrix node from output
5. Rows: Survived, Columns: $R-Survived
6. Go to Appearance tab and enable Percentage of Row option
7. Run stream to see output

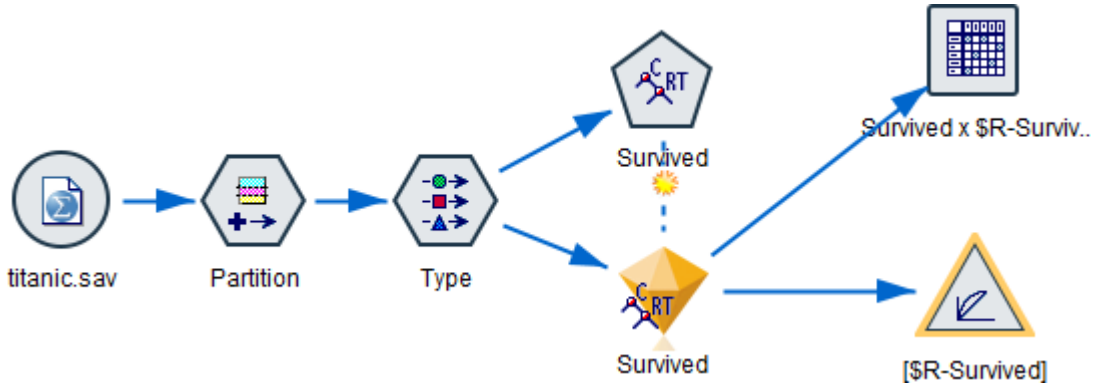


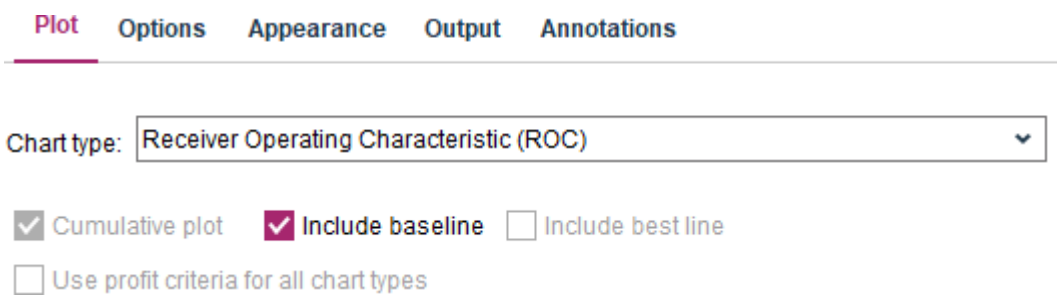


Bütün hüquqlar qorunur.

Building ROC curve

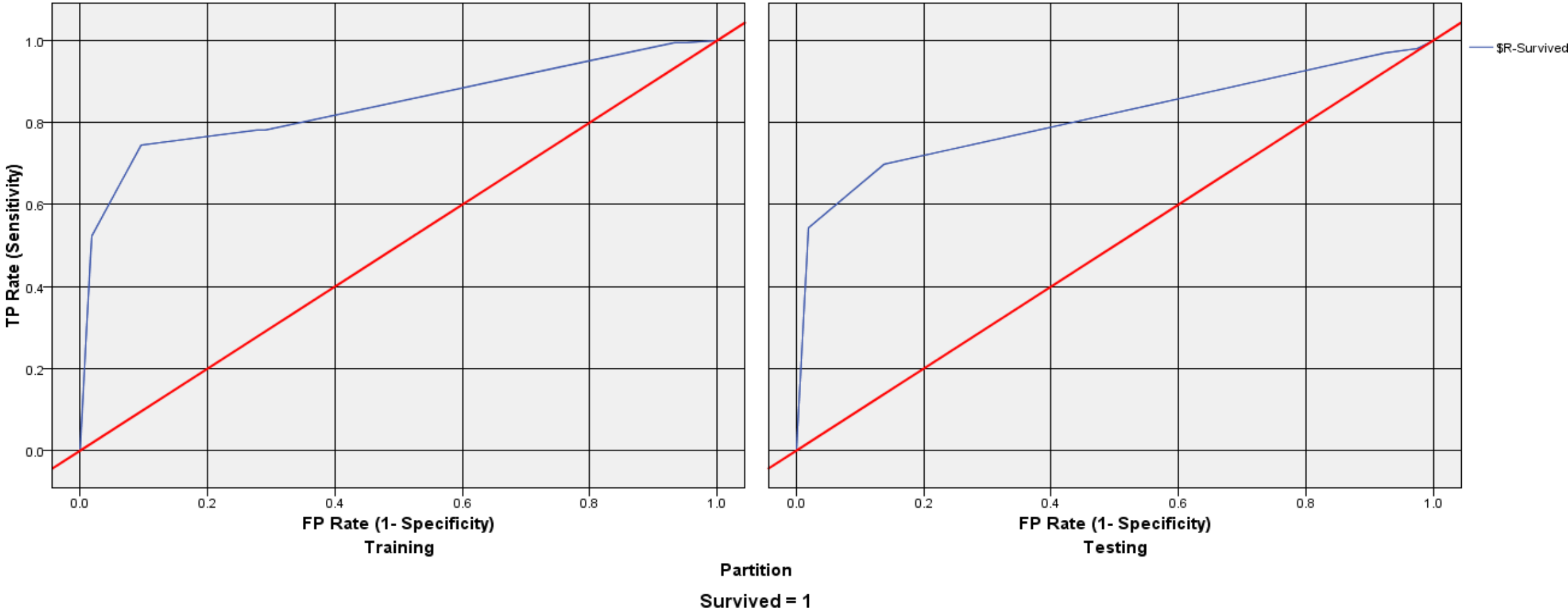
1. Open titanic partitioned.str stream file
2. Add **C&RT node** from Modeling palette
3. Run stream
4. Add matrix node from graphs
5. Choose ROC as Chart type
6. Run stream again





Bütün hüquqlar qorunur.

ROC curve



Bütün hüquqlar qorunur.

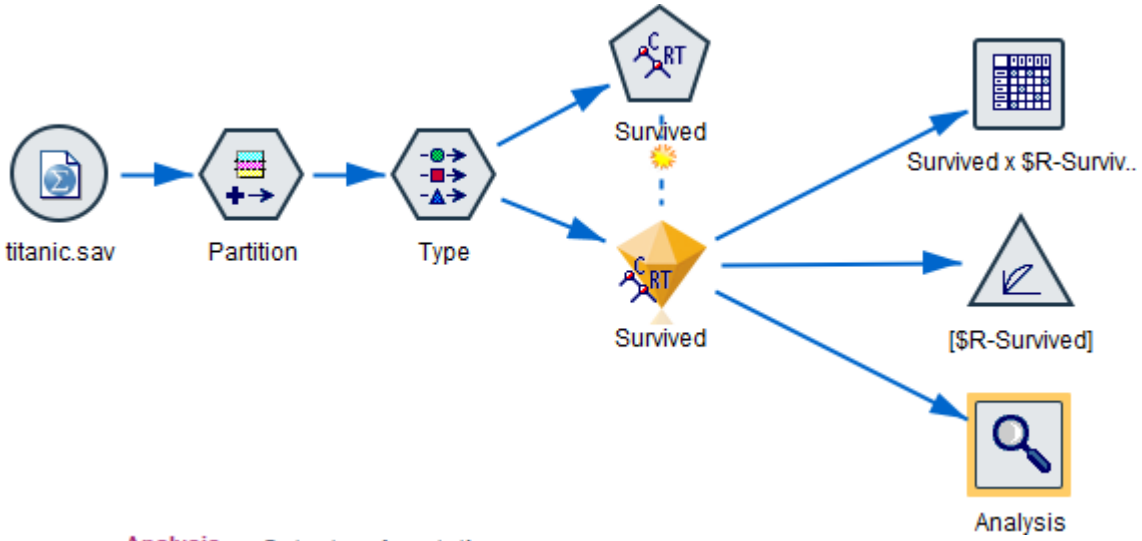
A Gini Coefficient

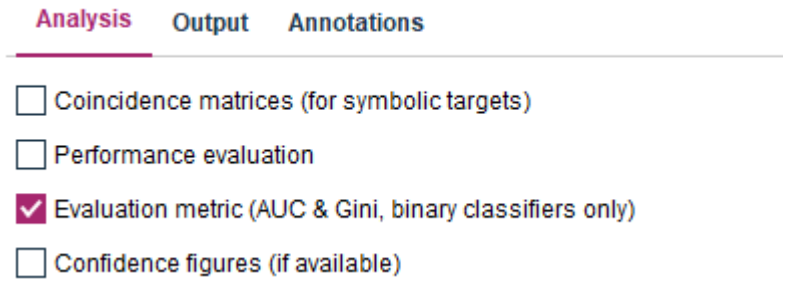
- A Gini coefficient can be used to evaluate the performance of a classifier. A classifier is a model that identifies to which class or category a request belongs to. In credit risk, classifiers can identify if an applicant belongs to the creditworthy or the uncreditworthy categories.
- A Gini is most commonly used for imbalanced datasets where the probability alone makes it difficult to predict an outcome. The Gini coefficient is a standard metric in risk assessment because the likelihood of default is relatively low. In the consumer finance industry, Gini can assess the accuracy of a prediction around whether a loan applicant will repay or default.
- Gini is measured in values between 0 and 1, where a score of 1 means that the model is 100% accurate in predicting the outcome. A score of 1 only exists in theory. In practice, the closer the Gini is to 1, the better. Whereas a Gini score equal to 0 means the model is entirely inaccurate. To achieve a score of 0, the model would have to ascribe random values to every prediction

Bütün hüquqlar qorunur.

AUC evaluation metric

1. Open titanic partitioned.str stream file
2. Add **C&RT node** from **Modeling** palette
3. Run stream
4. Add Analysis node from output
5. Choose Evaluation
6. Run stream again





Bütün hüquqlar qorunur.

AUC Metric for Train and Test set

Results for output field Survived

Individual Models

Comparing $R-Survived with Survived

'Partition'	1_Training		2_Testing	
Correct	530	84.39%	210	79.85%
Wrong	98	15.61%	53	20.15%
Total	628		263	

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-Survived	0.841	0.683	0.809	0.618

Bütün hüquqlar qorunur.

Thank you!

Data Science Academy