1 Data import and preparation

This part of the document deals with the data preparation of the provided cooper wire data before the data analysis.

1.1 Data import

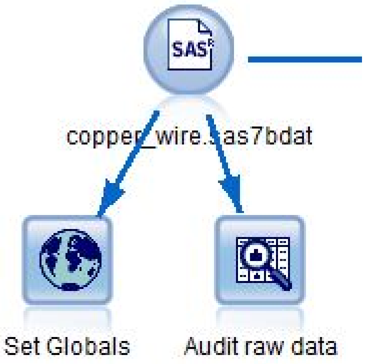
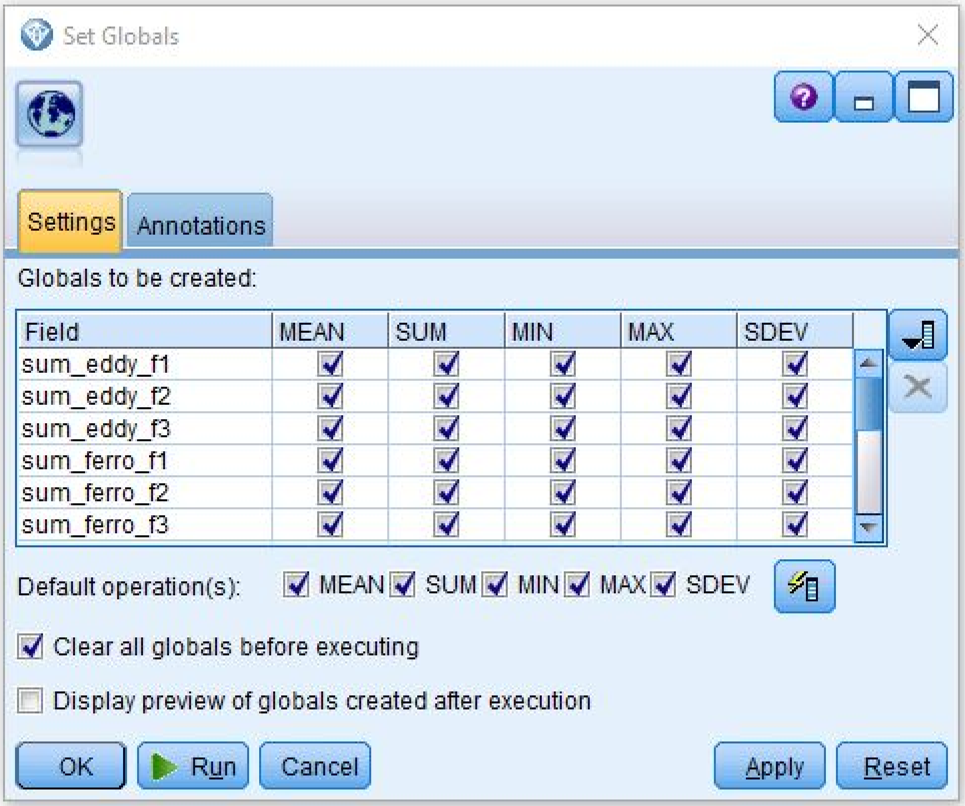


Abbildung 1: Data import in Stream

The data is imported via the node *SAS file*. The node *Set Globals* is used for setting the audited data results of the raw imported data as global values, which get used later on for the data preparation. The node *Data Audit* is used for analyzing the raw data.



Globals to be created:

Field	MEAN	SUM	MIN	MAX	SDEV
sum_eddy_f1	☑	☑	☑	☑	☑
sum_eddy_f2	☑	☑	☑	☑	☑
sum_eddy_f3	☑	☑	☑	☑	☑
sum_ferro_f1	☑	☑	☑	☑	☑
sum_ferro_f2	☑	☑	☑	☑	☑
sum_ferro_f3	☑	☑	☑	☑	☑

Default operation(s): ☑ MEAN ☑ SUM ☑ MIN ☑ MAX ☑ SDEV

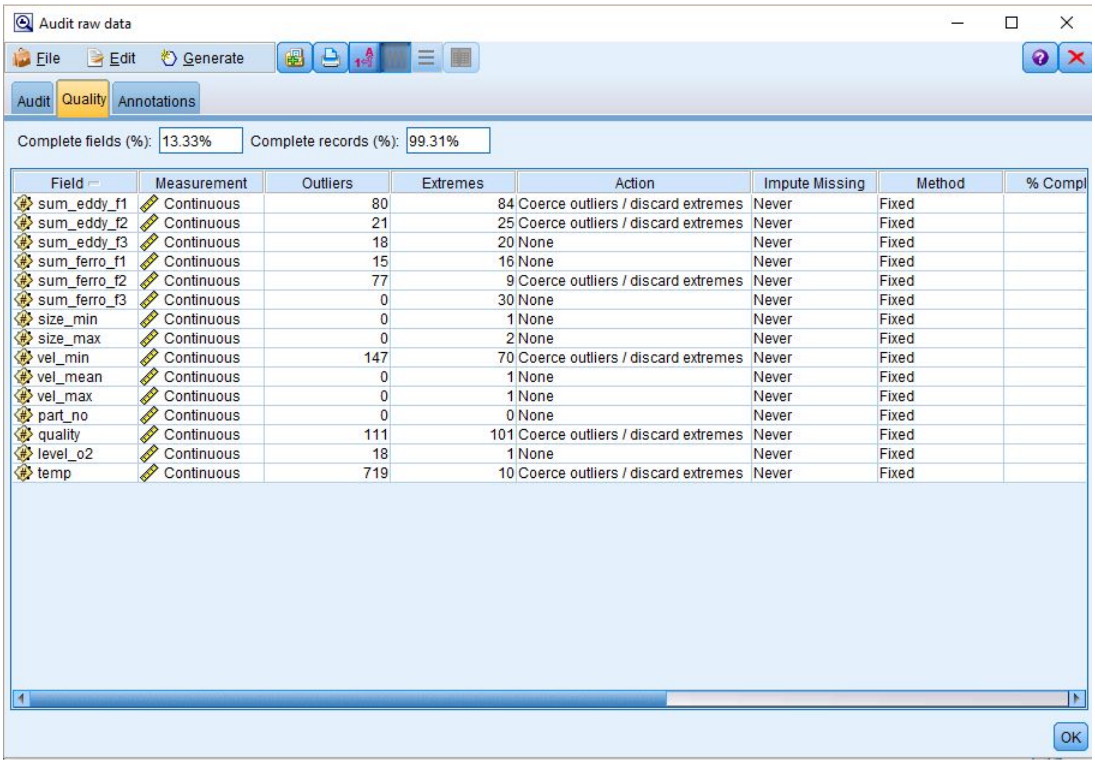
☑ Clear all globals before executing

□ Display preview of globals created after execution

Abbildung 2: Set audit results as global values in the stream

1.2 Data preparation

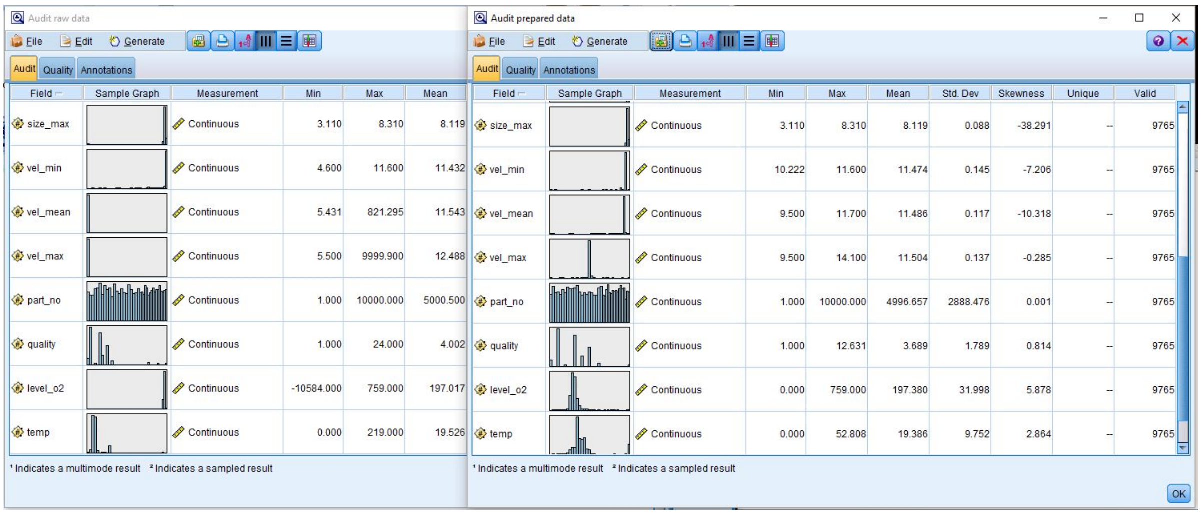
The outliers and extremes where determined during the audit of the raw data.



Complete fields (%): 13.33%   Complete records (%): 99.31%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Compl
sum_eddy_f1	Continuous	80	84	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f2	Continuous	21	25	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f3	Continuous	18	20	None	Never	Fixed	
sum_ferro_f1	Continuous	15	16	None	Never	Fixed	
sum_ferro_f2	Continuous	77	9	Coerce outliers / discard extremes	Never	Fixed	
sum_ferro_f3	Continuous	0	30	None	Never	Fixed	
size_min	Continuous	0	1	None	Never	Fixed	
size_max	Continuous	0	2	None	Never	Fixed	
vel_min	Continuous	147	70	Coerce outliers / discard extremes	Never	Fixed	
vel_mean	Continuous	0	1	None	Never	Fixed	
vel_max	Continuous	0	1	None	Never	Fixed	
part_no	Continuous	0	0	None	Never	Fixed	
quality	Continuous	111	101	Coerce outliers / discard extremes	Never	Fixed	
level_o2	Continuous	18	1	None	Never	Fixed	
temp	Continuous	719	10	Coerce outliers / discard extremes	Never	Fixed	

Abbildung 3: Audit of the raw data



Field	Sample Graph	Measurement	Min	Max	Mean
size_max		Continuous	3.110	8.310	8.119
vel_min		Continuous	4.600	11.600	11.432
vel_mean		Continuous	5.431	821.295	11.543
vel_max		Continuous	5.500	9999.900	12.488
part_no		Continuous	1.000	10000.000	5000.500
quality		Continuous	1.000	24.000	4.002
level_o2		Continuous	-10584.000	759.000	197.017
temp		Continuous	0.000	219.000	19.526

[1] Indicates a multimode result [2] Indicates a sampled result

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
size_max		Continuous	3.110	8.310	8.119	0.088	-38.291	--	9765
vel_min		Continuous	10.222	11.600	11.474	0.145	-7.206	--	9765
vel_mean		Continuous	9.500	11.700	11.486	0.117	-10.318	--	9765
vel_max		Continuous	9.500	14.100	11.504	0.137	-0.285	--	9765
part_no		Continuous	1.000	10000.000	4996.657	2888.476	0.001	--	9765
quality		Continuous	1.000	12.631	3.689	1.789	0.814	--	9765
level_o2		Continuous	0.000	759.000	197.380	31.998	5.878	--	9765
temp		Continuous	0.000	52.808	19.386	9.752	2.864	--	9765

[1] Indicates a multimode result [2] Indicates a sampled result

Abbildung 4: Audit of the raw data

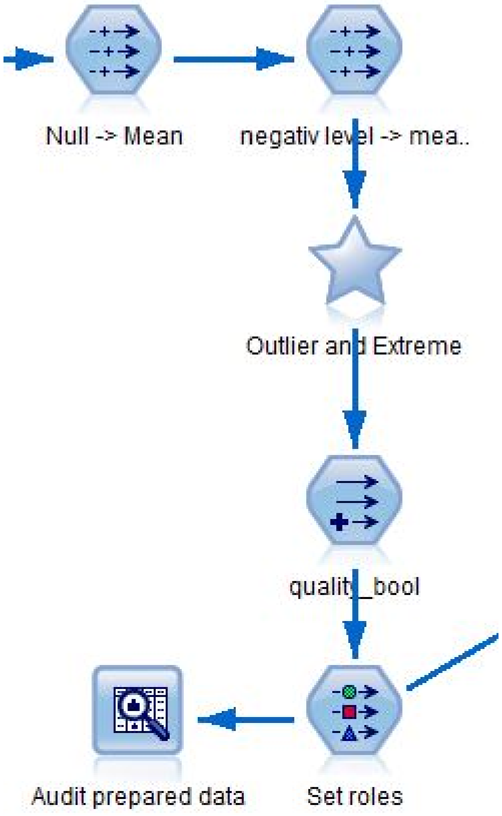


Abbildung 5: Flow of data preparation tasks

This flow prepares the data for the later analysis. The following tasks are performed:

- Null values will be replaced with the mean value set by the *Set Global* node
- The negative value of the field *temp* will be replaced with the global mean of this field
- The outliers and extremes will be handled as you can see in image 3
- A new filed will be created *quality_bool* which represents the quality state good or false
- The fields which are not considered to be relevant will be set as ignored and the field *quality_bool* will be set as the target field for the further analysis

1.3 Predictive Model

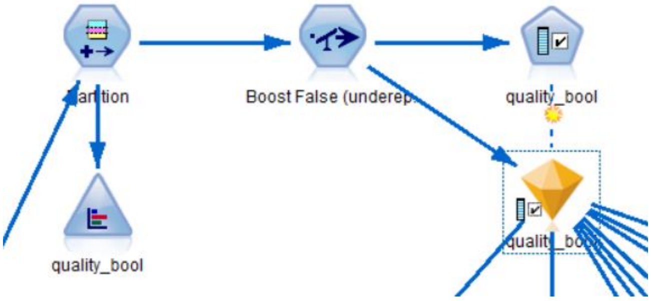
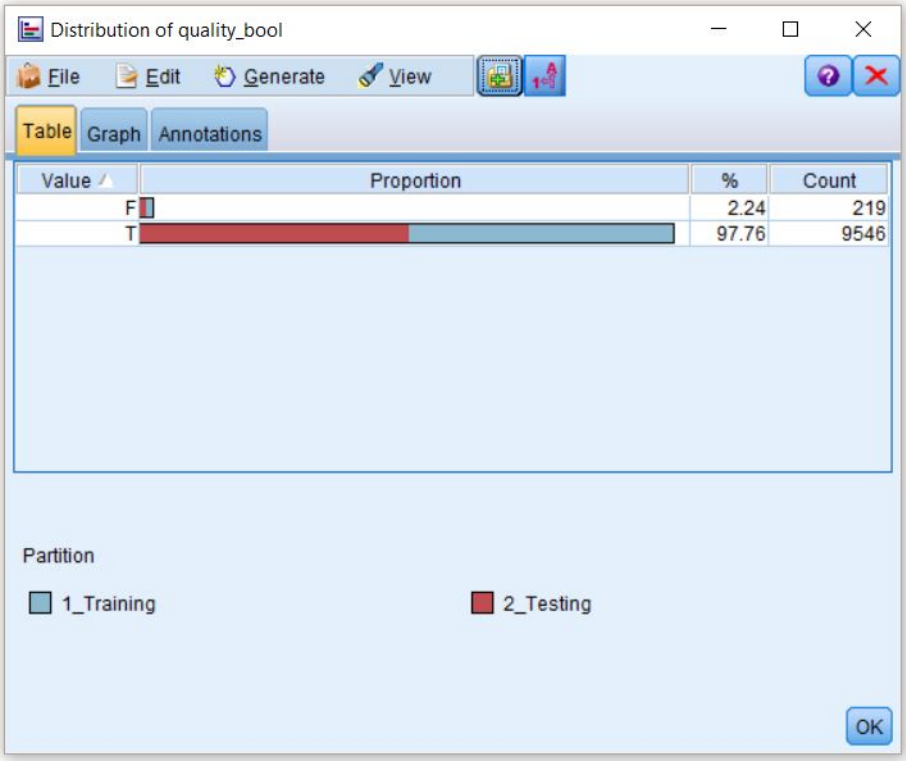


Abbildung 6: Flow of further data preparation

This part of the stream prepares the data in the following way:

- The node *Partition* splits the data in a training and testing data.
- The node *Distribution* shows us that the field *quality_bool* is very bad distributed. *(False=2.24, True=97.76)*
-



Value	Proportion	%	Count
F		2.24	219
T		97.76	9546

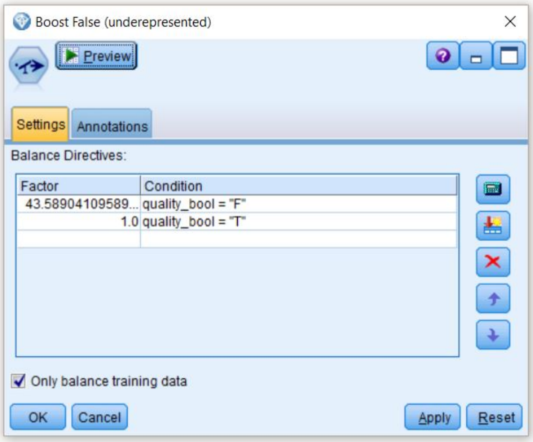
Partition

🟦 1_Training

🟥 2_Testing

Abbildung 7: Badly distributed *quality_bool*

As we can see that the *False* quality is underrepresented compared to the *True* quality.



Balance Directives:

Factor	Condition
43.58904109589...	quality_bool = "F"
1.0	quality_bool = "T"

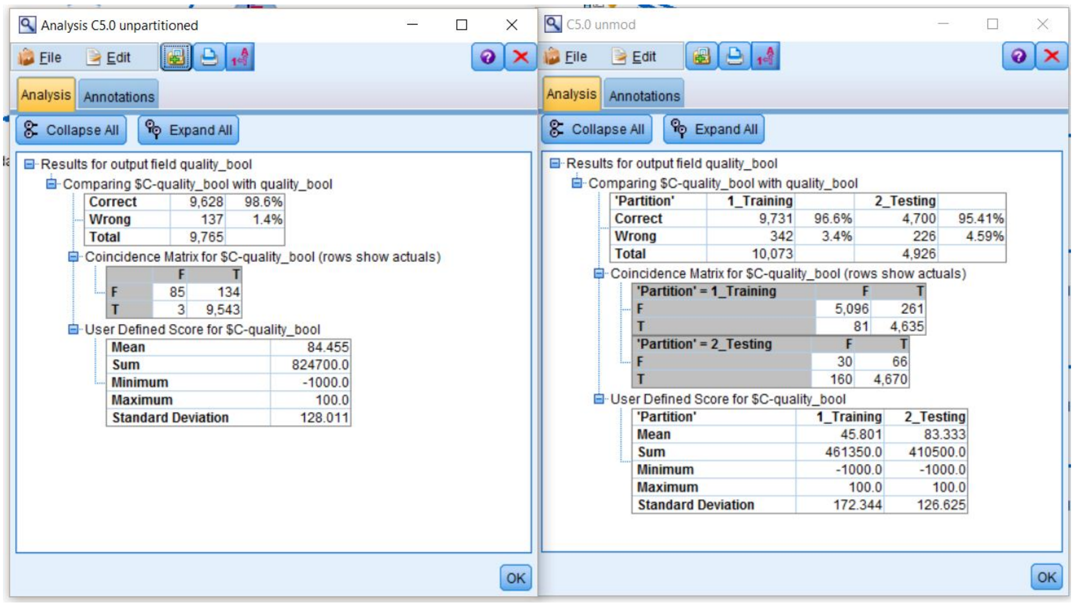
☑ Only balance training data

Abbildung 8: Boost of quality *False*

The node *Balance* has been generated by the node *Distribution* and boost the representation of the *False* quality. After this nodes follows the node *Field selection* which removes fields which are not related to the *target*.

The node *Field selection* has reduced the count of fields from *15* down to *10*, therefore has removed *5* fields.

1.3.1 Not partitioned data to feature selection and boost



Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

Correct	9,628	98.6%
Wrong	137	1.4%
Total	9,765	

Coincidence Matrix for $C-quality_bool (rows show actuals)

	F	T
F	85	134
T	3	9,543

User Defined Score for $C-quality_bool

Mean	84.455
Sum	824700.0
Minimum	-1000.0
Maximum	100.0
Standard Deviation	128.011

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,731	96.6%	4,700	95.41%
Wrong	342	3.4%	226	4.59%
Total	10,073		4,926	

Coincidence Matrix for $C-quality_bool (rows show actuals)

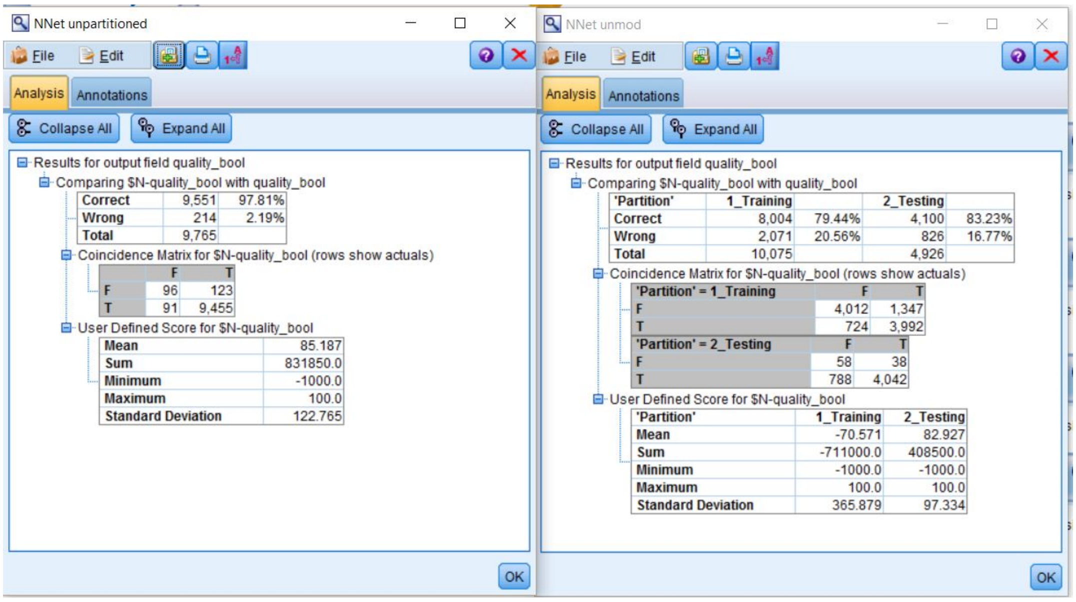
'Partition' = 1_Training	F	T
F	5,096	261
T	81	4,635

'Partition' = 2_Testing	F	T
F	30	66
T	160	4,670

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	45.801	83.333
Sum	461350.0	410500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	172.344	126.625

Abbildung 9: C5.0 with no partitioned and partitioned data



NNet unpartitioned

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

Correct	9,551	97.81%
Wrong	214	2.19%
Total	9,765	

Coincidence Matrix for $N-quality_bool (rows show actuals)

	F	T
F	96	123
T	91	9,455

User Defined Score for $N-quality_bool

Mean	85.187
Sum	831850.0
Minimum	-1000.0
Maximum	100.0
Standard Deviation	122.765

NNet unmod

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,004	79.44%	4,100	83.23%
Wrong	2,071	20.56%	826	16.77%
Total	10,075		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

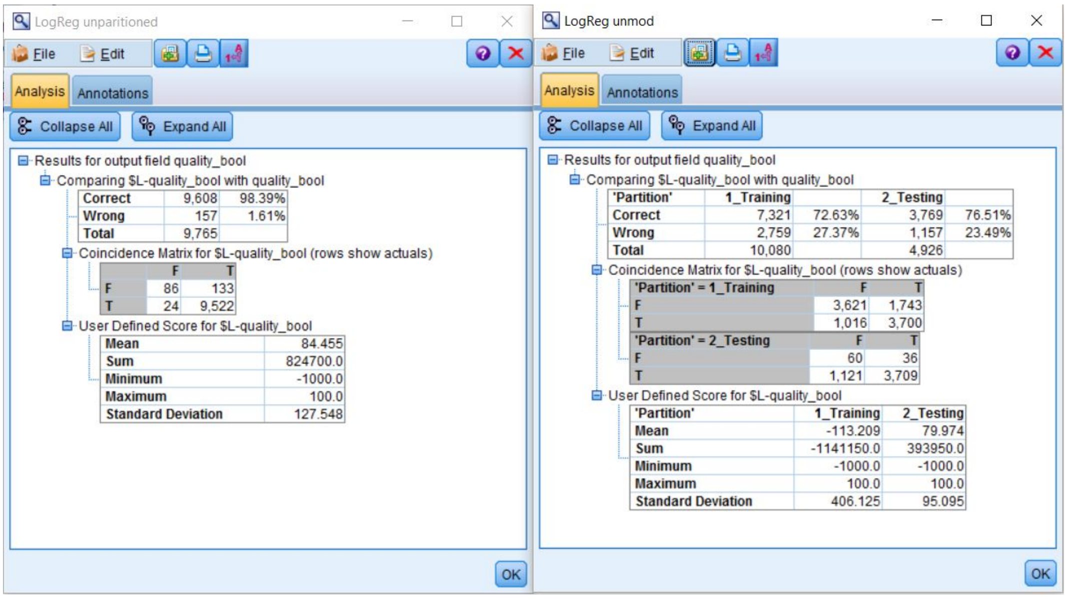
'Partition' = 1_Training	F	T
F	4,012	1,347
T	724	3,992

'Partition' = 2_Testing	F	T
F	58	38
T	788	4,042

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-70.571	82.927
Sum	-711000.0	408500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	365.879	97.334

Abbildung 10: Neuronal Net with no partitioned and partitioned data



LogReg unparitioned

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

Correct	9,608	98.39%
Wrong	157	1.61%
Total	9,765	

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

	F	T
F	86	133
T	24	9,522

User Defined Score for $L-quality_bool

Mean	84.455
Sum	824700.0
Minimum	-1000.0
Maximum	100.0
Standard Deviation	127.548

LogReg unmod

Results for output field quality_bool

Comparing $L-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	7,321	72.63%	3,769	76.51%
Wrong	2,759	27.37%	1,157	23.49%
Total	10,080		4,926	

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

'Partition' = 1_Training	F	T
F	3,621	1,743
T	1,016	3,700

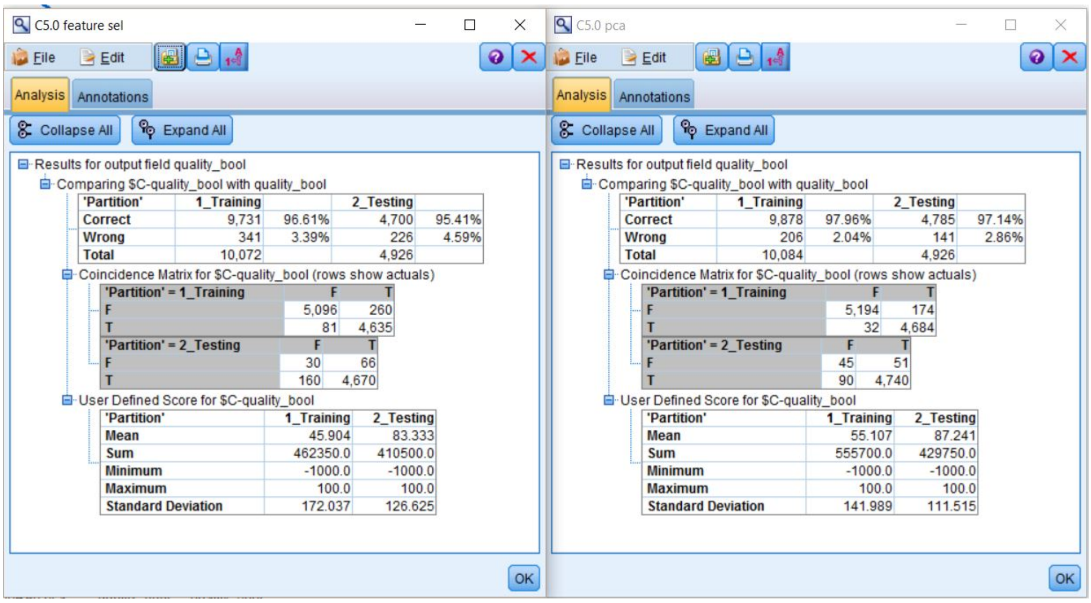
'Partition' = 2_Testing	F	T
F	60	36
T	1,121	3,709

User Defined Score for $L-quality_bool

'Partition'	1_Training	2_Testing
Mean	-113.209	79.974
Sum	-1141150.0	393950.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	406.125	95.095

Abbildung 11: Logistic Regression with no partitioned and partitioned data

1.3.2 Feature selection to PCA



C5.0 feature sel

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,731	96.61%	4,700	95.41%
Wrong	341	3.39%	226	4.59%
Total	10,072		4,926	

Coincidence Matrix for $C-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	5,096	260
T	81	4,635

'Partition' = 2_Testing	F	T
F	30	66
T	160	4,670

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	45.904	83.333
Sum	462350.0	410500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	172.037	126.625

C5.0 pca

Results for output field quality_bool

Comparing $C-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	9,878	97.96%	4,785	97.14%
Wrong	206	2.04%	141	2.86%
Total	10,084		4,926	

Coincidence Matrix for $C-quality_bool (rows show actuals)

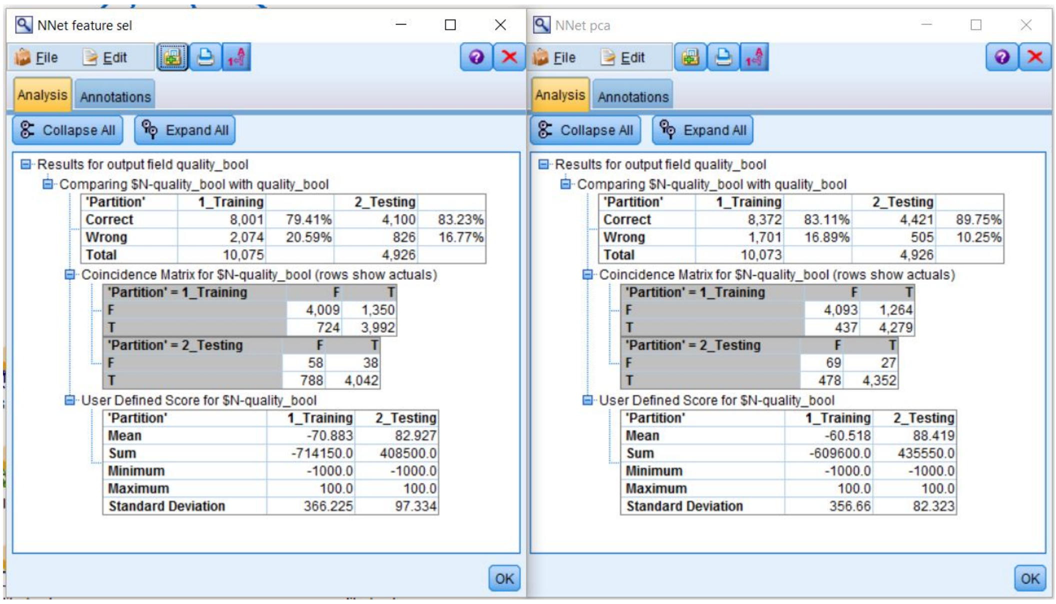
'Partition' = 1_Training	F	T
F	5,194	174
T	32	4,684

'Partition' = 2_Testing	F	T
F	45	51
T	90	4,740

User Defined Score for $C-quality_bool

'Partition'	1_Training	2_Testing
Mean	55.107	87.241
Sum	555700.0	429750.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	141.989	111.515

Abbildung 12: C5.0 feature selection to PCA



NNet feature sel

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,001	79.41%	4,100	83.23%
Wrong	2,074	20.59%	826	16.77%
Total	10,075		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	4,009	1,350
T	724	3,992

'Partition' = 2_Testing	F	T
F	58	38
T	788	4,042

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-70.883	82.927
Sum	-714150.0	408500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	366.225	97.334

NNet pca

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,372	83.11%	4,421	89.75%
Wrong	1,701	16.89%	505	10.25%
Total	10,073		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

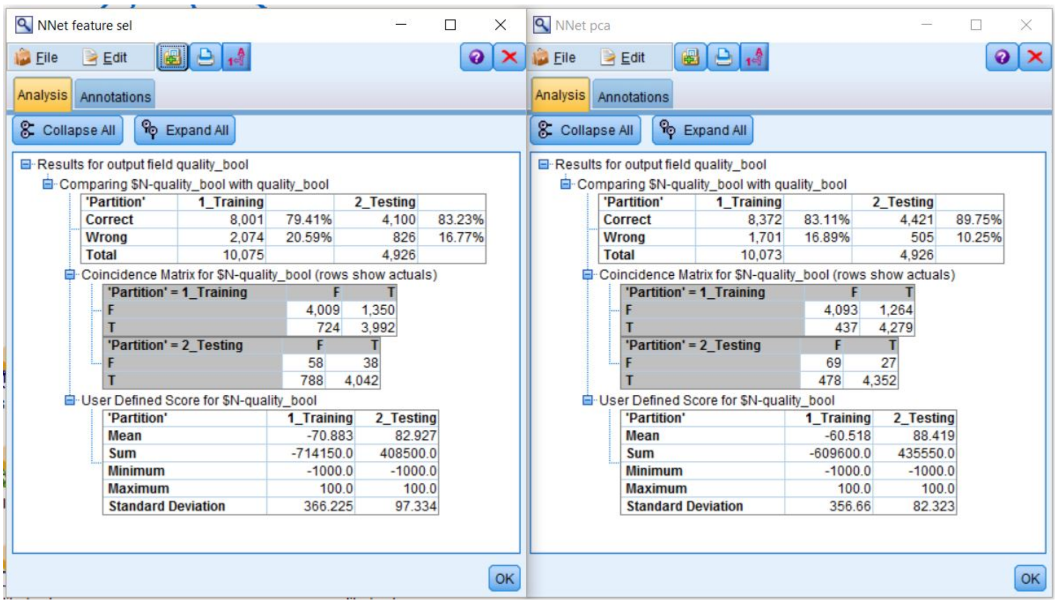
'Partition' = 1_Training	F	T
F	4,093	1,264
T	437	4,279

'Partition' = 2_Testing	F	T
F	69	27
T	478	4,352

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-60.518	88.419
Sum	-609600.0	435550.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	356.66	82.323

Abbildung 13: Neuronal Net feature selection to PCA



Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,001	79.41%	4,100	83.23%
Wrong	2,074	20.59%	826	16.77%
Total	10,075		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	4,009	1,350
T	724	3,992

'Partition' = 2_Testing	F	T
F	58	38
T	788	4,042

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-70.883	82.927
Sum	-714150.0	408500.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	366.225	97.334

Results for output field quality_bool

Comparing $N-quality_bool with quality_bool

'Partition'	1_Training		2_Testing	
Correct	8,372	83.11%	4,421	89.75%
Wrong	1,701	16.89%	505	10.25%
Total	10,073		4,926	

Coincidence Matrix for $N-quality_bool (rows show actuals)

'Partition' = 1_Training	F	T
F	4,093	1,264
T	437	4,279

'Partition' = 2_Testing	F	T
F	69	27
T	478	4,352

User Defined Score for $N-quality_bool

'Partition'	1_Training	2_Testing
Mean	-60.518	88.419
Sum	-609600.0	435550.0
Minimum	-1000.0	-1000.0
Maximum	100.0	100.0
Standard Deviation	356.66	82.323

Abbildung 14: Logistic Regression feature selection to PCA

1.3.3 Fine tuning of PCA part