به نام خدا

محمد سعید حیدری (۴۰۰۴۲۲۰۷۵)

تحلیل دیتاست boston house price در ibm spss modeler

درس داده کاوی : دکتر پرند

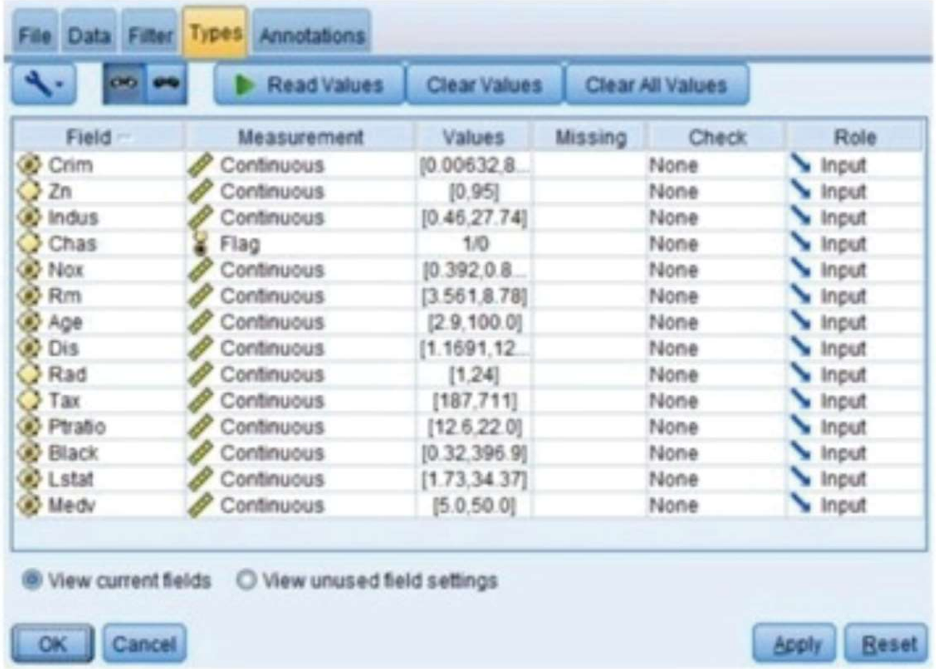
دانشگاه شهید بهشتی

پیاده سازی بر اساس گام های CRISP

Cross-Industry Standard Process for Data Mining (CRISP-DM)

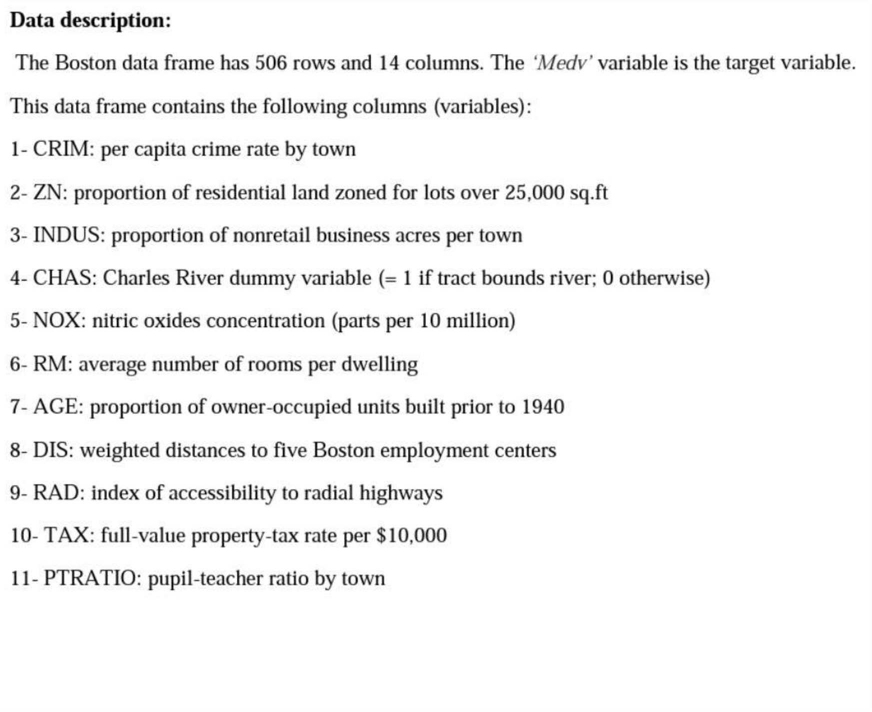
1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

1. Loading Data



Field	Measurement	Values	Missing	Check	Role
Crim	Continuous	[0.00632,8…		None	Input
Zn	Continuous	[0,95]		None	Input
Indus	Continuous	[0.46,27.74]		None	Input
Chas	Flag	1/0		None	Input
Nox	Continuous	[0.392,0.8…		None	Input
Rm	Continuous	[3.561,8.78]		None	Input
Age	Continuous	[2.9,100.0]		None	Input
Dis	Continuous	[1.1691,12…		None	Input
Rad	Continuous	[1,24]		None	Input
Tax	Continuous	[187,711]		None	Input
Ptratio	Continuous	[12.6,22.0]		None	Input
Black	Continuous	[0.32,396.9]		None	Input
Lstat	Continuous	[1.73,34.37]		None	Input
Medv	Continuous	[5.0,50.0]		None	Input

● View current fields ○ View unused field settings



Data description:

The Boston data frame has 506 rows and 14 columns. The *'Medv'* variable is the target variable.

This data frame contains the following columns (variables):

1- CRIM: per capita crime rate by town

2- ZN: proportion of residential land zoned for lots over 25,000 sq.ft

3- INDUS: proportion of nonretail business acres per town

4- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5- NOX: nitric oxides concentration (parts per 10 million)

6- RM: average number of rooms per dwelling

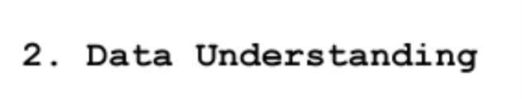
7- AGE: proportion of owner-occupied units built prior to 1940

8- DIS: weighted distances to five Boston employment centers

9- RAD: index of accessibility to radial highways

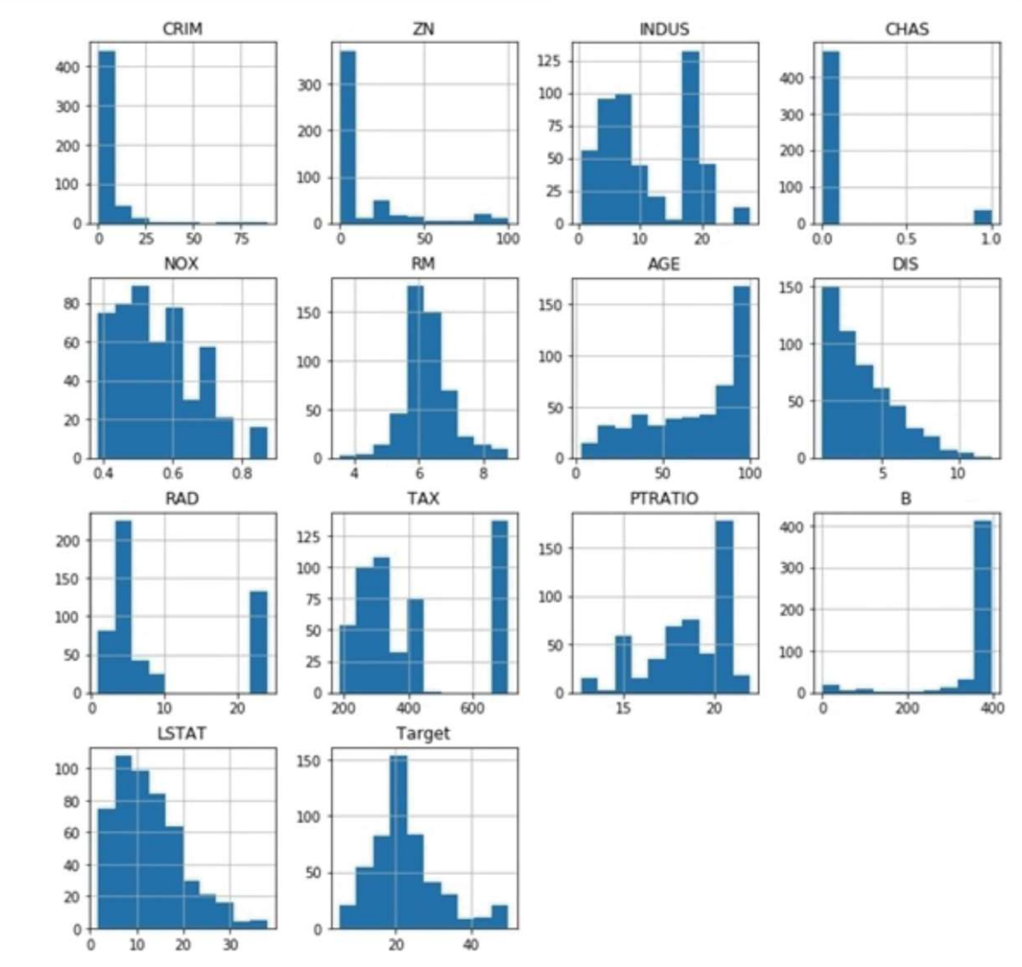
10- TAX: full-value property-tax rate per $10,000

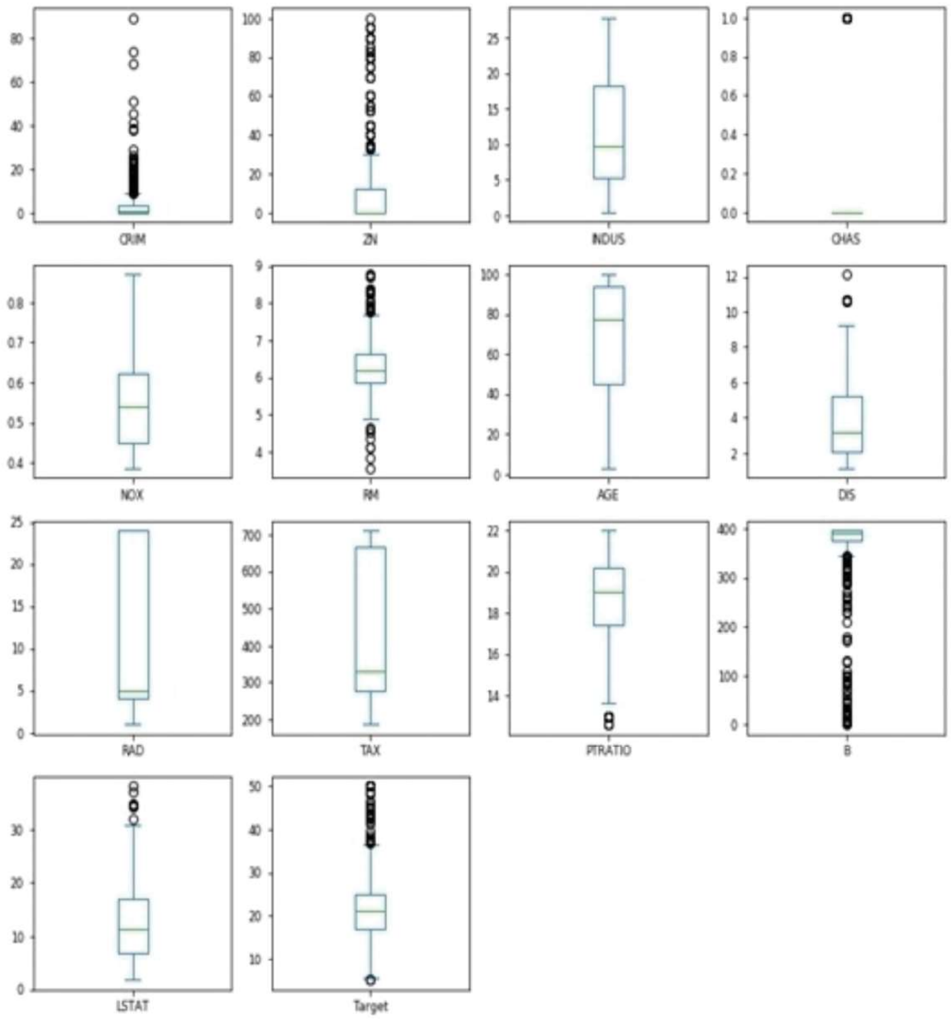
11- PTRATIO: pupil-teacher ratio by town



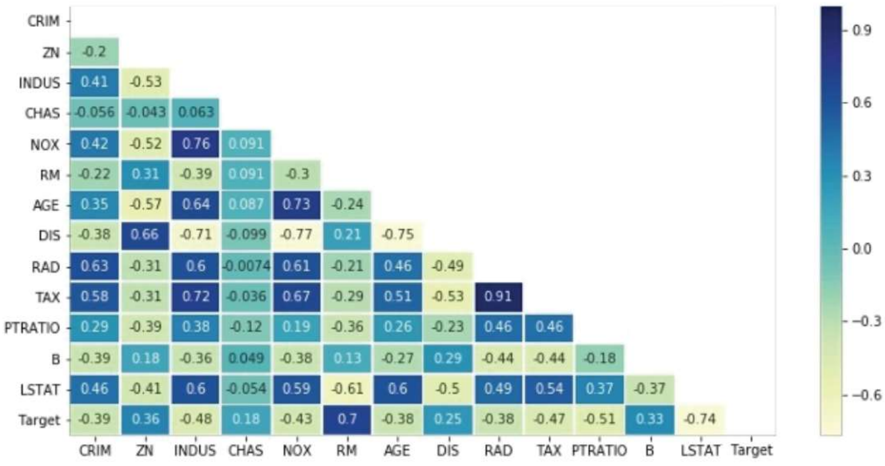
2. Data Understanding

به منظور کشف نقاط پرت و توزیع داده ها، می توانیم از ابزارهای رسم نمودار مانند Boxplots و هیستوگرام استفاده کنیم.

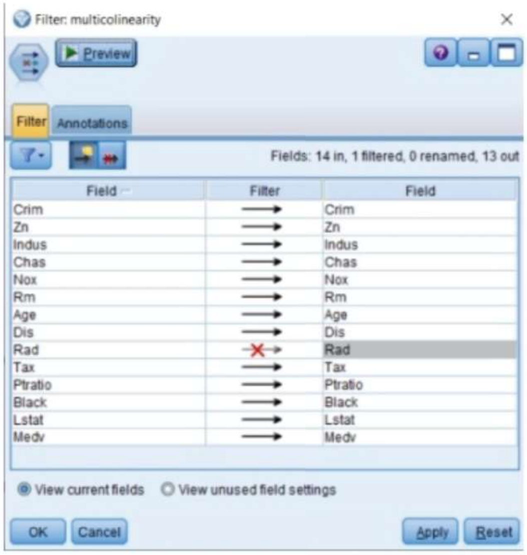




3. Data Preparation



به دلیل اینکه TAX و RAD دارای correlation= 0.91 یکی از آنها را انتخاب و دیگری را خط میزنیم

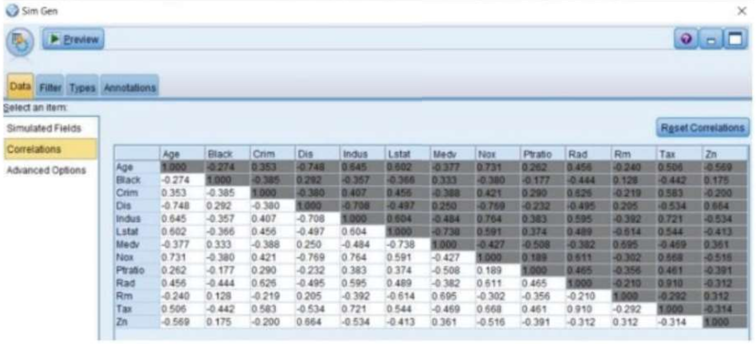


Filter: multicolinearity

Fields: 14 in, 1 filtered, 0 renamed, 13 out

Field	Filter	Field
Crim	→	Crim
Zn	→	Zn
Indus	→	Indus
Chas	→	Chas
Nox	→	Nox
Rm	→	Rm
Age	→	Age
Dis	→	Dis
Rad	✗	Rad
Tax	→	Tax
Ptratio	→	Ptratio
Black	→	Black
Lstat	→	Lstat
Medv	→	Medv

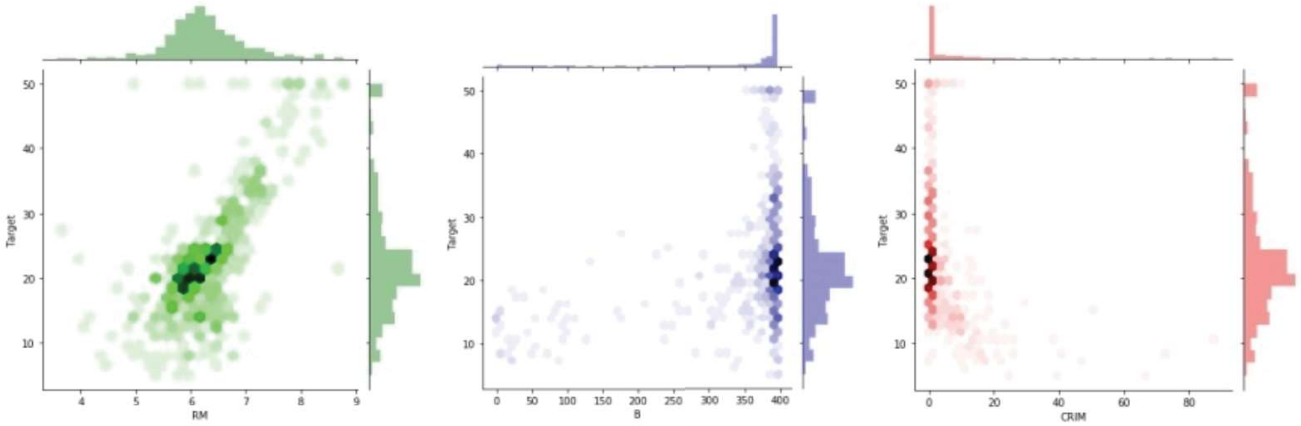
● View current fields ○ View unused field settings



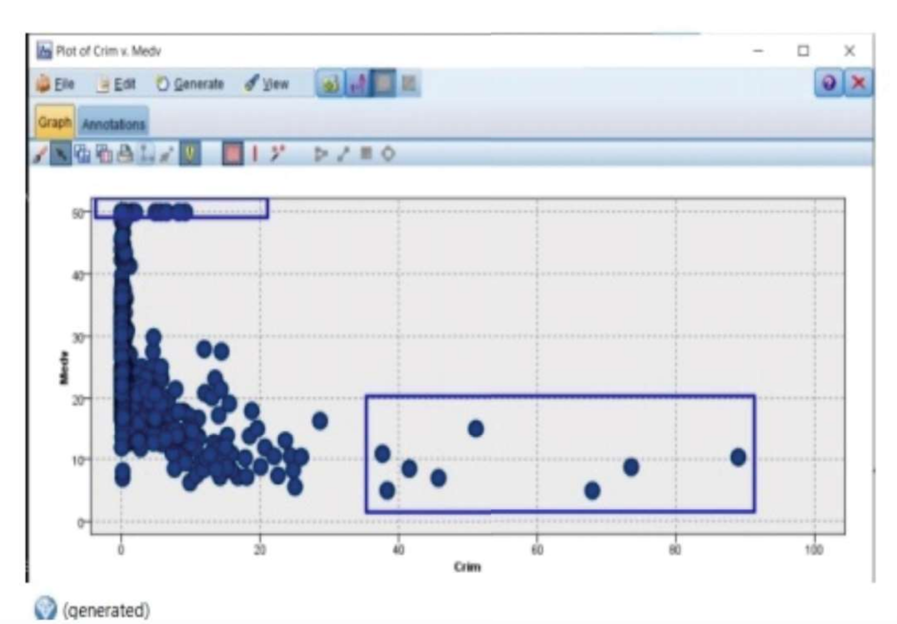
Sim Gen

	Age	Black	Crim	Dis	Indus	Lstat	Medv	Nox	Ptratio	Rad	Rm	Tax	Zn
Age	1.000	-0.274	0.353	-0.748	0.645	0.602	-0.377	0.731	0.262	0.456	-0.240	0.506	-0.569
Black	-0.274	1.000	-0.385	0.292	-0.357	-0.366	0.333	-0.380	-0.177	-0.444	0.128	-0.442	0.175
Crim	0.353	-0.385	1.000	-0.380	0.407	0.456	-0.388	0.421	0.290	0.626	-0.219	0.583	-0.200
Dis	-0.748	0.292	-0.380	1.000	-0.708	-0.497	0.250	-0.769	-0.232	-0.495	0.205	-0.534	0.664
Indus	0.645	-0.357	0.407	-0.708	1.000	0.604	-0.484	0.764	0.383	0.595	-0.392	0.721	-0.534
Lstat	0.602	-0.366	0.456	-0.497	0.604	1.000	-0.738	0.591	0.374	0.489	-0.614	0.544	-0.413
Medv	-0.377	0.333	-0.388	0.250	-0.484	-0.738	1.000	-0.427	-0.508	-0.382	0.695	-0.469	0.361
Nox	0.731	-0.380	0.421	-0.769	0.764	0.591	-0.427	1.000	0.189	0.611	-0.302	0.668	-0.516
Ptratio	0.262	-0.177	0.290	-0.232	0.383	0.374	-0.508	0.189	1.000	0.465	-0.356	0.461	-0.391
Rad	0.456	-0.444	0.626	-0.495	0.595	0.489	-0.382	0.611	0.465	1.000	-0.210	0.910	-0.312
Rm	-0.240	0.128	-0.219	0.205	-0.392	-0.614	0.695	-0.302	-0.356	-0.210	1.000	-0.292	0.312
Tax	0.506	-0.442	0.583	-0.534	0.721	0.544	-0.469	0.668	0.461	0.910	-0.292	1.000	-0.314
Zn	-0.569	0.175	-0.200	0.664	-0.534	-0.413	0.361	-0.516	-0.391	-0.312	0.312	-0.314	1.000

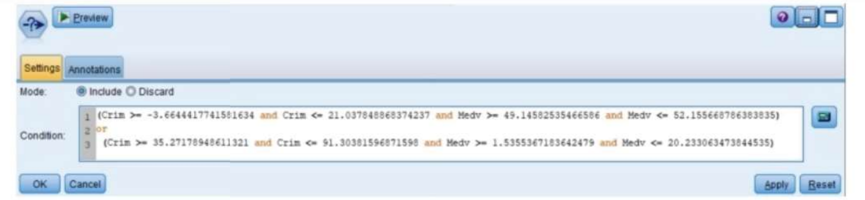
When we visualize the data, we see that the data seems to be capped at 50. The data points with a *'Medv'* value of 50 are likely contain censored or missing values. We nullify these points by using Interactions option in Plot View and Select Node or by using Filler Node.



مشاهده distribution داده و بررسی outliers ها



استفاده از گزینه Interactions در Plot View

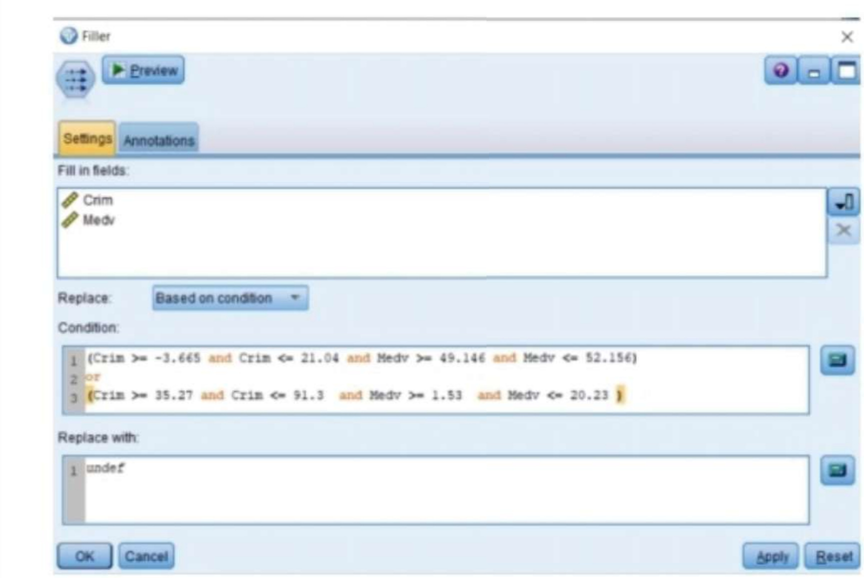


Mode: ● Include ○ Discard

Condition:

```
(Crim >= -3.6644417741581634 and Crim <= 21.037848868374237 and Medv >= 49.14582535466586 and Medv <= 52.155668786383835)
or
 (Crim >= 35.27178948611321 and Crim <= 91.30381596871598 and Medv >= 1.5355367183642479 and Medv <= 20.233063473844535)
```

استفاده از گزینه nullify برای missing data



Fill in fields:

- Crim
- Medv

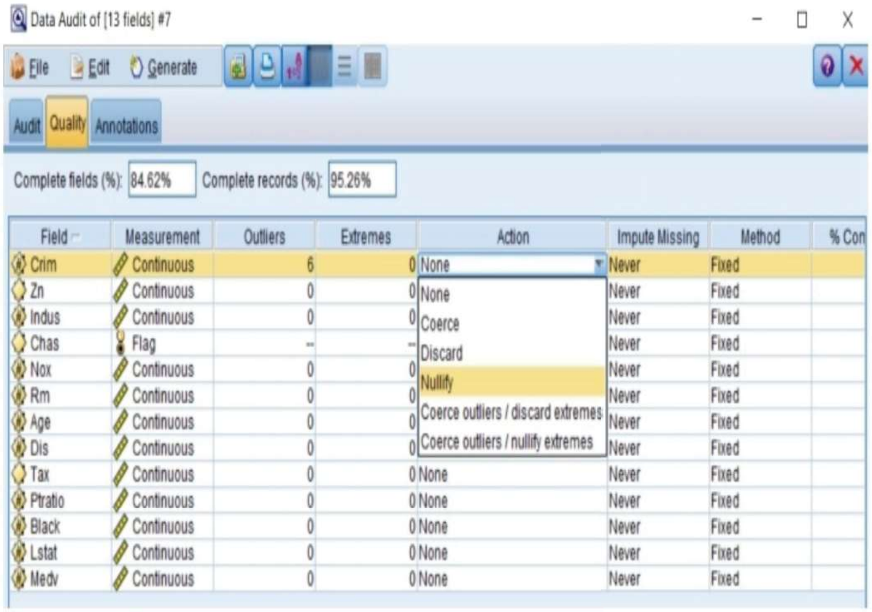
Replace: Based on condition

Condition:

```
(Crim >= -3.665 and Crim <= 21.04 and Medv >= 49.146 and Medv <= 52.156)
or
(Crim >= 35.27 and Crim <= 91.3  and Medv >= 1.53  and Medv <= 20.23 )
```

Replace with:

```
undef
```



Data Audit of [13 fields] #7

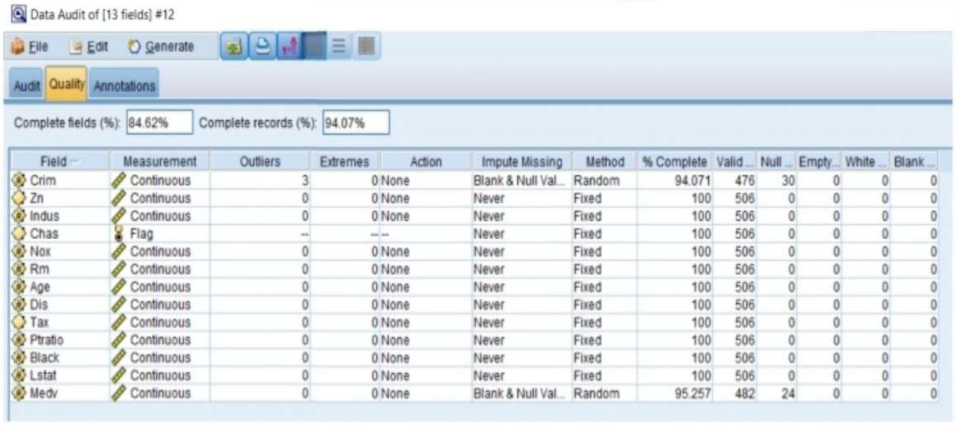
File
Edit
Generate
Audit
Quality
Annotations
None
Coerce
Discard
Nullify
Coerce outliers / discard extremes
Coerce outliers / nullify extremes

Complete fields (%): 84.62% Complete records (%): 95.26%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Con
Crim	Continuous	6	0	None	Never	Fixed	
Zn	Continuous	0	0		Never	Fixed	
Indus	Continuous	0	0		Never	Fixed	
Chas	Flag	--	--		Never	Fixed	
Nox	Continuous	0	0		Never	Fixed	
Rm	Continuous	0	0		Never	Fixed	
Age	Continuous	0	0		Never	Fixed	
Dis	Continuous	0	0		Never	Fixed	
Tax	Continuous	0	0	None	Never	Fixed	
Ptratio	Continuous	0	0	None	Never	Fixed	
Black	Continuous	0	0	None	Never	Fixed	
Lstat	Continuous	0	0	None	Never	Fixed	
Medv	Continuous	0	0	None	Never	Fixed	

Handling Outliers

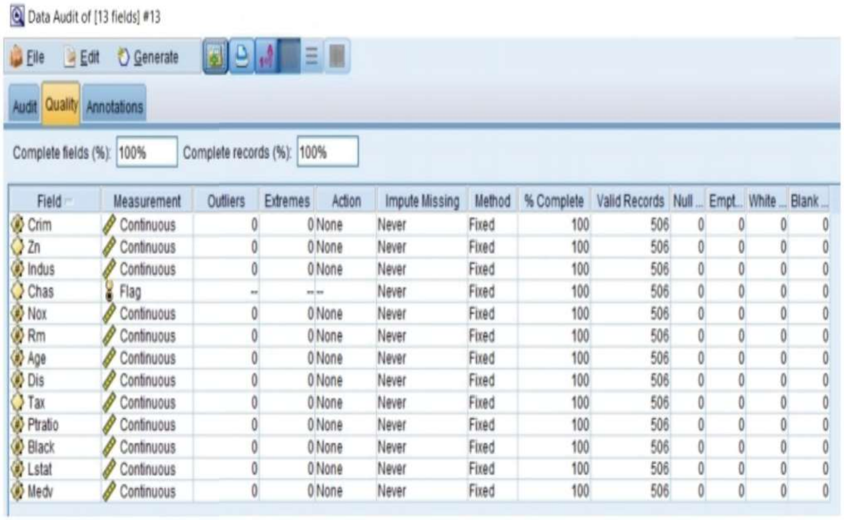
سپس از روش‌های مختلف داده miss را پر می‌کنیم



Data Audit of [13 fields] #12

Complete fields (%): 84.62% Complete records (%): 94.07%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid ...	Null ...	Empty...	White ...	Blank ...
Crim	Continuous	3	0	None	Blank & Null Val...	Random	94.071	476	30	0	0	0
Zn	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Indus	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Chas	Flag	--	--	--	Never	Fixed	100	506	0	0	0	0
Nox	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Rm	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Age	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Dis	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Tax	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Ptratio	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Black	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Lstat	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Medv	Continuous	0	0	None	Blank & Null Val...	Random	95.257	482	24	0	0	0



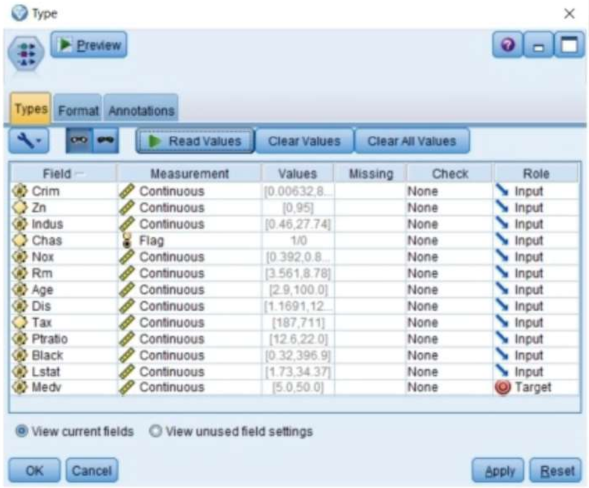
Data Audit of [13 fields] #13

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null ...	Empt...	White ...	Blank ...
Crim	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Zn	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Indus	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Chas	Flag	--	--	--	Never	Fixed	100	506	0	0	0	0
Nox	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Rm	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Age	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Dis	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Tax	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Ptratio	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Black	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Lstat	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0
Medv	Continuous	0	0	None	Never	Fixed	100	506	0	0	0	0

پایان data cleaning

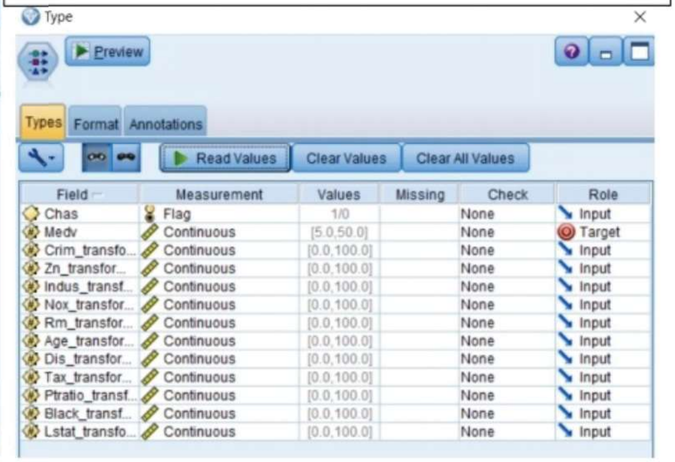
سپس از روش MimMax داده ها را نرمالایز می‌کنیم



Field	Measurement	Values	Missing	Check	Role
Crim	Continuous	[0.00632,8...		None	Input
Zn	Continuous	[0,95]		None	Input
Indus	Continuous	[0.46,27.74]		None	Input
Chas	Flag	1/0		None	Input
Nox	Continuous	[0.392,0.8...		None	Input
Rm	Continuous	[3.561,8.78]		None	Input
Age	Continuous	[2.9,100.0]		None	Input
Dis	Continuous	[1.1691,12...		None	Input
Tax	Continuous	[187,711]		None	Input
Ptratio	Continuous	[12.6,22.0]		None	Input
Black	Continuous	[0.32,396.9]		None	Input
Lstat	Continuous	[1.73,34.37]		None	Input
Medv	Continuous	[5.0,50.0]		None	Target

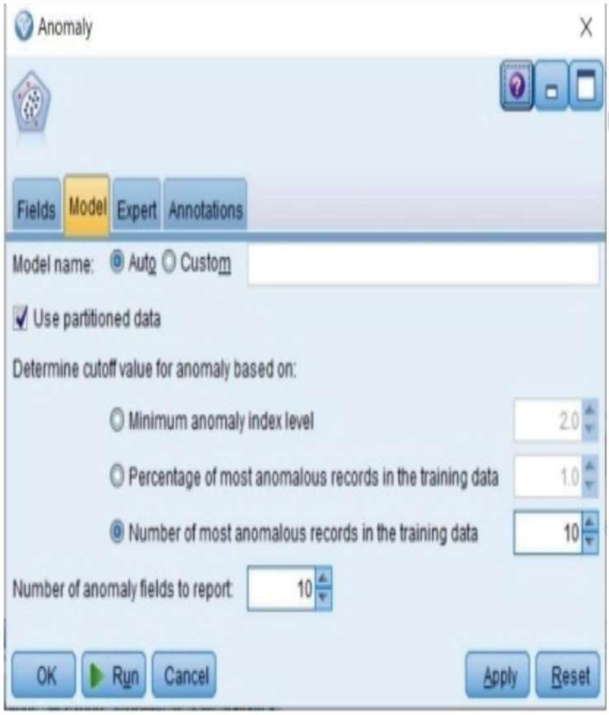
● View current fields ○ View unused field settings

MinMax Scaling Data: using Auto Data Prep Node



Field	Measurement	Values	Missing	Check	Role
Chas	Flag	1/0		None	Input
Medv	Continuous	[5.0,50.0]		None	Target
Crim_transfo...	Continuous	[0.0,100.0]		None	Input
Zn_transfor...	Continuous	[0.0,100.0]		None	Input
Indus_transf...	Continuous	[0.0,100.0]		None	Input
Nox_transfor...	Continuous	[0.0,100.0]		None	Input
Rm_transfor...	Continuous	[0.0,100.0]		None	Input
Age_transfor...	Continuous	[0.0,100.0]		None	Input
Dis_transfor...	Continuous	[0.0,100.0]		None	Input
Tax_transfor...	Continuous	[0.0,100.0]		None	Input
Ptratio_transf...	Continuous	[0.0,100.0]		None	Input
Black_transf...	Continuous	[0.0,100.0]		None	Input
Lstat_transfo...	Continuous	[0.0,100.0]		None	Input

سپس بررسی نویز و آنومالی



Model name: ● Auto ○ Custom

☑ Use partitioned data

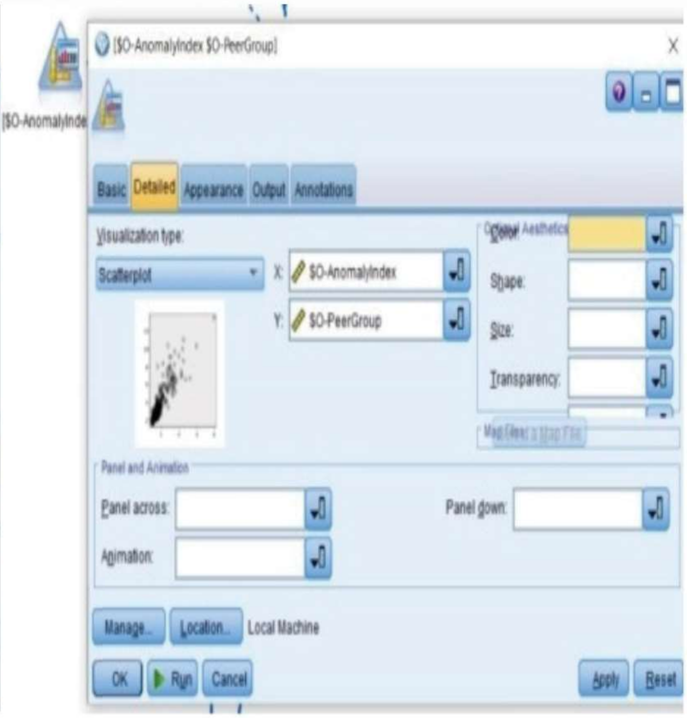
Determine cutoff value for anomaly based on:

○ Minimum anomaly index level 2.0

○ Percentage of most anomalous records in the training data 1.0

● Number of most anomalous records in the training data 10

Number of anomaly fields to report: 10



Visualization type: Scatterplot

X: $0-AnomalyIndex

Y: $0-PeerGroup

Color:

Shape:

Size:

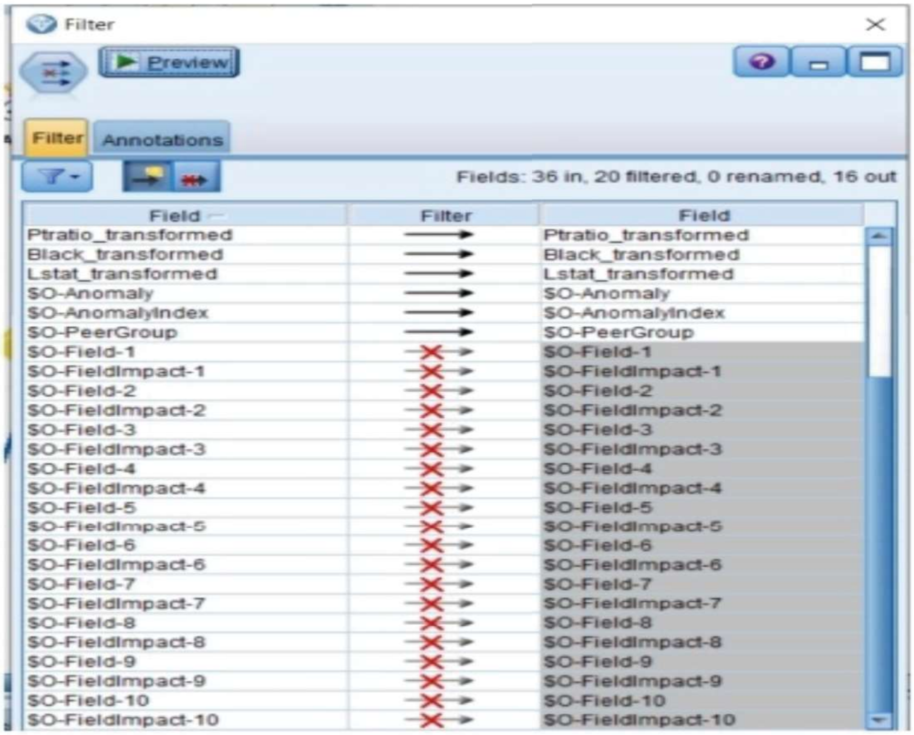
Transparency:

Panel across:

Panel down:

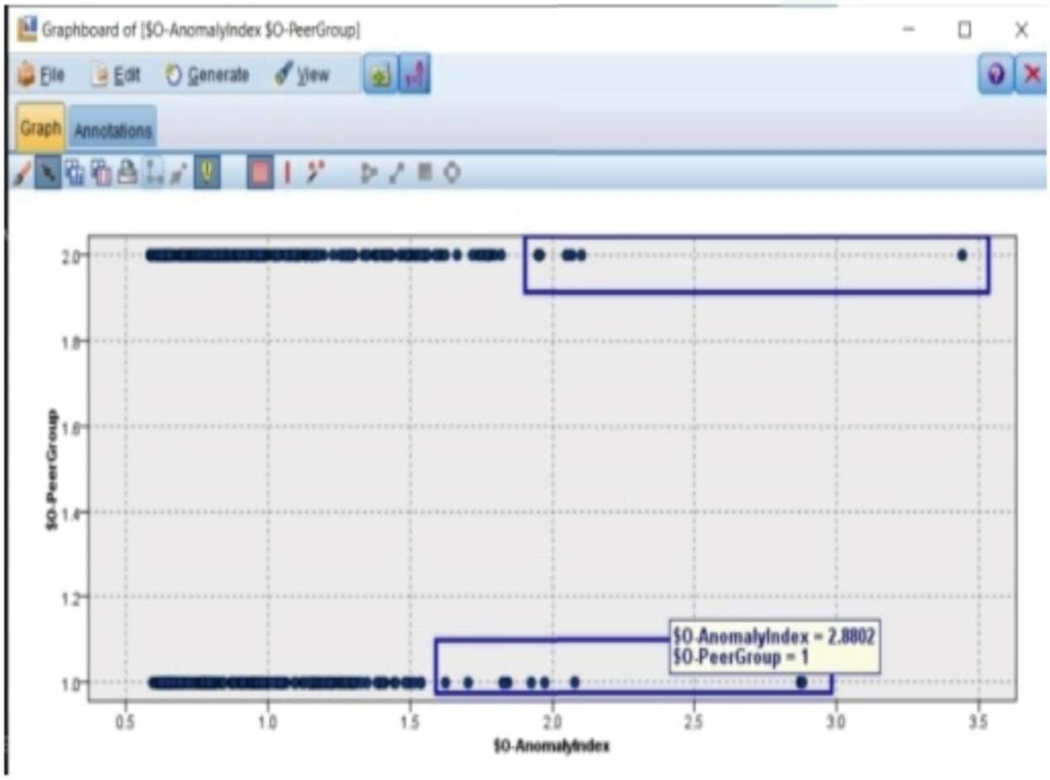
Animation:

Local Machine



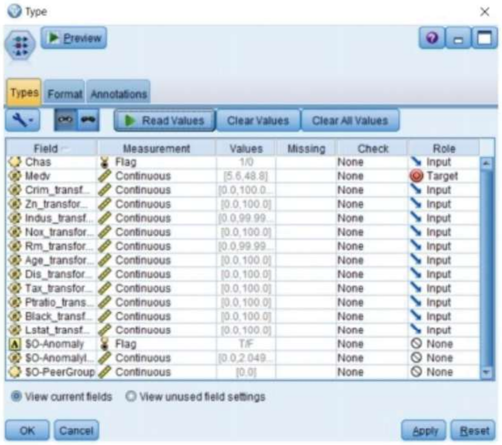
Fields: 36 in, 20 filtered, 0 renamed, 16 out

Field	Filter	Field
Ptratio_transformed	→	Ptratio_transformed
Black_transformed	→	Black_transformed
Lstat_transformed	→	Lstat_transformed
$O-Anomaly	→	$O-Anomaly
$O-AnomalyIndex	→	$O-AnomalyIndex
$O-PeerGroup	→	$O-PeerGroup
$O-Field-1	✗→	$O-Field-1
$O-FieldImpact-1	✗→	$O-FieldImpact-1
$O-Field-2	✗→	$O-Field-2
$O-FieldImpact-2	✗→	$O-FieldImpact-2
$O-Field-3	✗→	$O-Field-3
$O-FieldImpact-3	✗→	$O-FieldImpact-3
$O-Field-4	✗→	$O-Field-4
$O-FieldImpact-4	✗→	$O-FieldImpact-4
$O-Field-5	✗→	$O-Field-5
$O-FieldImpact-5	✗→	$O-FieldImpact-5
$O-Field-6	✗→	$O-Field-6
$O-FieldImpact-6	✗→	$O-FieldImpact-6
$O-Field-7	✗→	$O-Field-7
$O-FieldImpact-7	✗→	$O-FieldImpact-7
$O-Field-8	✗→	$O-Field-8
$O-FieldImpact-8	✗→	$O-FieldImpact-8
$O-Field-9	✗→	$O-Field-9
$O-FieldImpact-9	✗→	$O-FieldImpact-9
$O-Field-10	✗→	$O-Field-10
$O-FieldImpact-10	✗→	$O-FieldImpact-10



که در اینجا نویز درون داده به وضوح مشخص است. آنرا رفع میکنیم

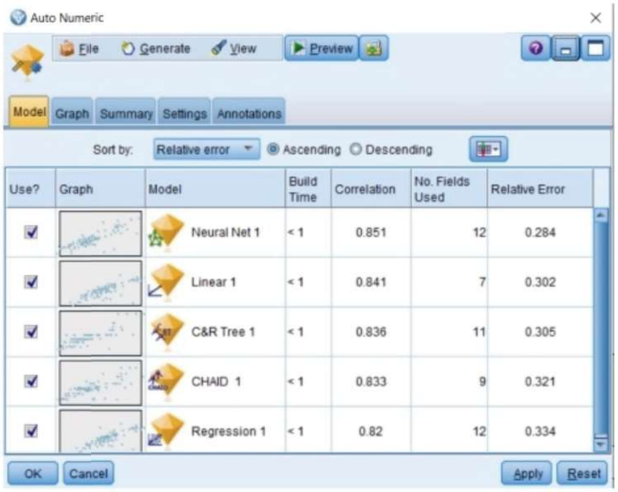
4. Modelling



Type

Field	Measurement	Values	Missing	Check	Role
Chas	Flag	1/0		None	Input
Medv	Continuous	[5.6,48.8]		None	Target
Crim_transf...	Continuous	[0.0,100.0...		None	Input
Zn_transfor...	Continuous	[0.0,100.0]		None	Input
Indus_transf...	Continuous	[0.0,99.99...		None	Input
Nox_transfor...	Continuous	[0.0,100.0]		None	Input
Rm_transfor...	Continuous	[0.0,99.99...		None	Input
Age_transfor...	Continuous	[0.0,100.0]		None	Input
Dis_transfor...	Continuous	[0.0,100.0]		None	Input
Tax_transfor...	Continuous	[0.0,100.0]		None	Input
Ptratio_trans...	Continuous	[0.0,100.0]		None	Input
Black_transf...	Continuous	[0.0,100.0]		None	Input
Lstat_transf...	Continuous	[0.0,100.0]		None	Input
$O-Anomaly	Flag	T/F		None	None
$O-Anomalyl...	Continuous	[0.0,2.049...		None	None
$O-PeerGroup	Continuous	[0.0]		None	None

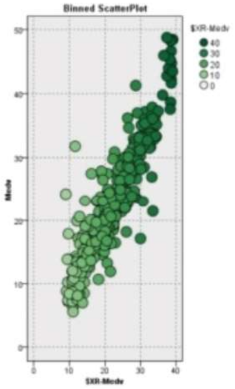
● View current fields ○ View unused field settings

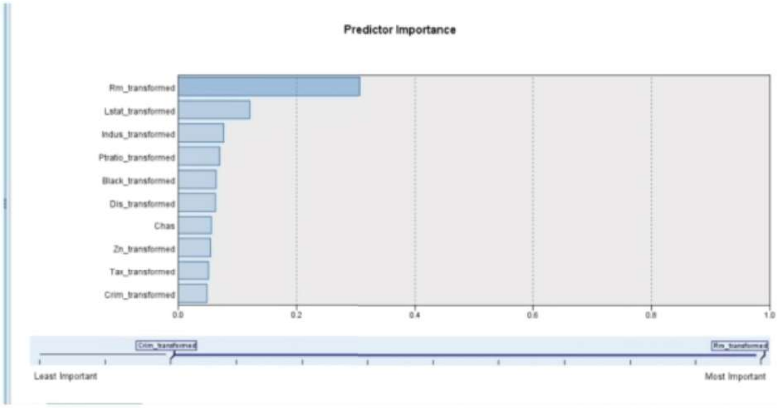


Auto Numeric

Sort by: Relative error ● Ascending ○ Descending

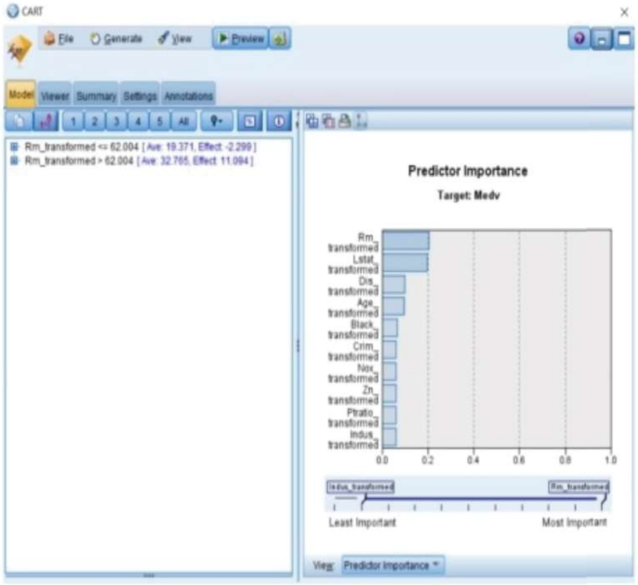
Use?	Graph	Model	Build Time	Correlation	No. Fields Used	Relative Error
☑		Neural Net 1	< 1	0.851	12	0.284
☑		Linear 1	< 1	0.841	7	0.302
☑		C&R Tree 1	< 1	0.836	11	0.305
☑		CHAID 1	< 1	0.833	9	0.321
☑		Regression 1	< 1	0.82	12	0.334





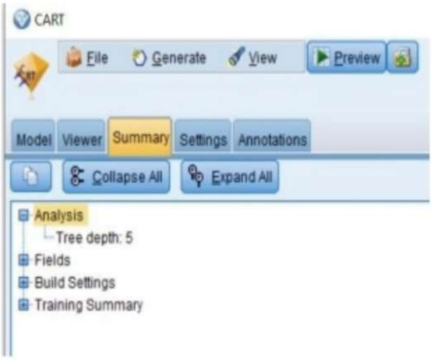
تک تک روش‌ها و الگوریتم ها رو امتحان می‌کنیم

CART:



CART

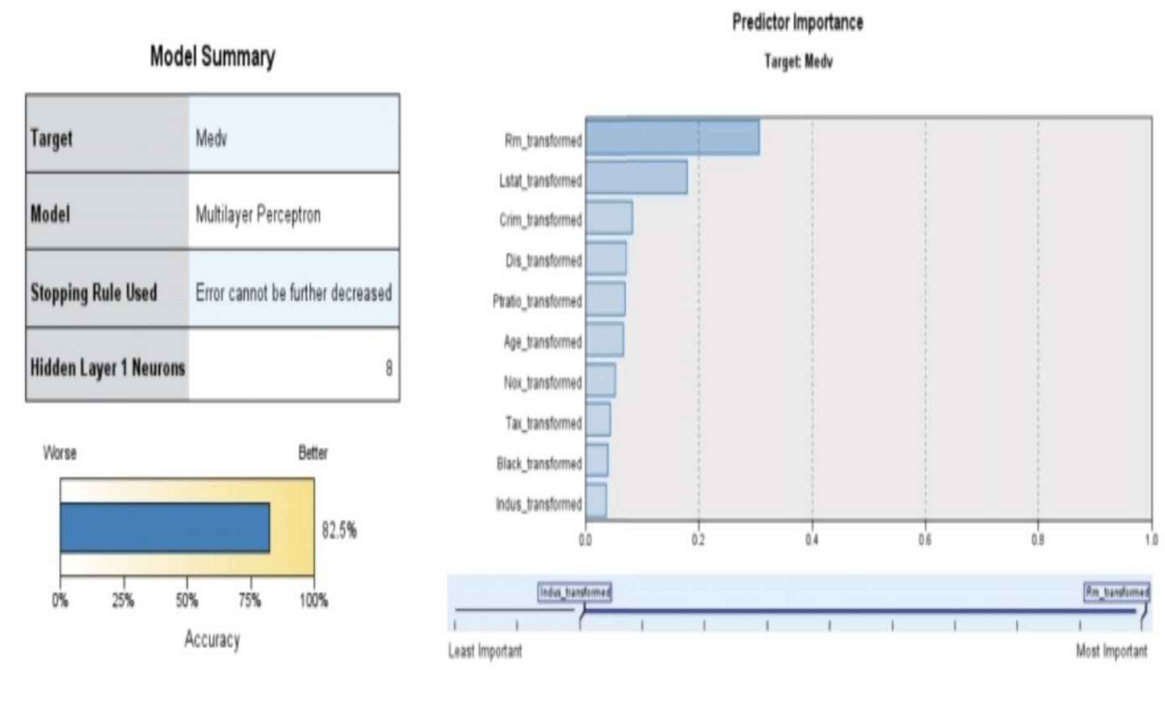
Rm_transformed <= 62.004 [Ave: 19.371, Effect: -2.299]
Rm_transformed > 62.004 [Ave: 32.765, Effect: 11.094]



CART

Analysis
Tree depth: 5
Fields
Build Settings
Training Summary

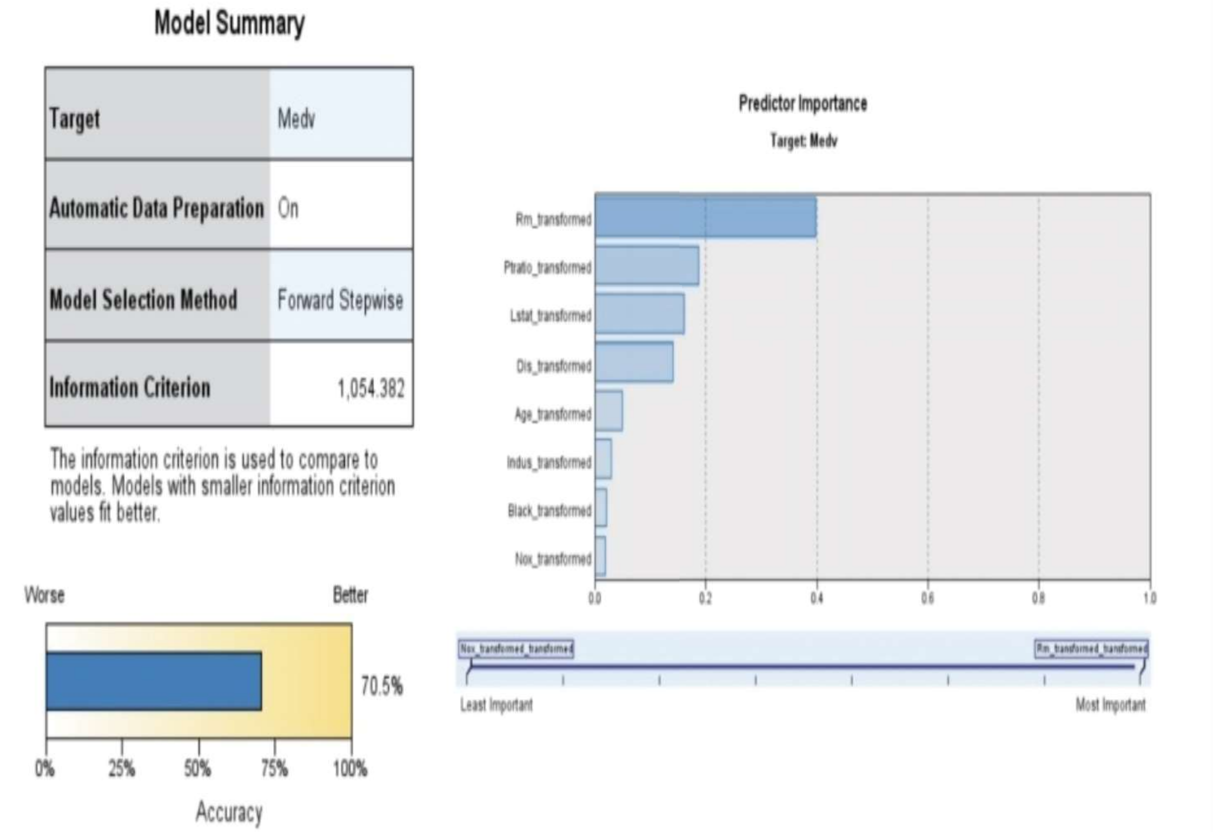
Neural Network:



Model Summary

Target	Medv
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	8

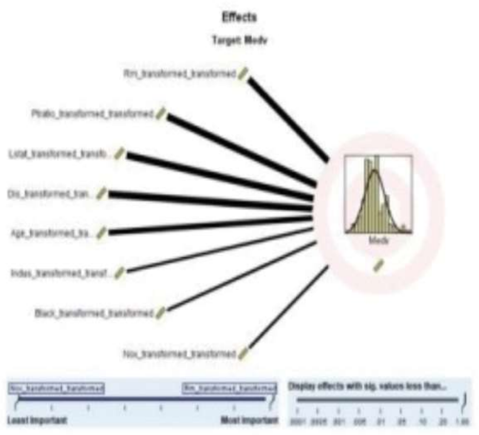
Linear Regression:

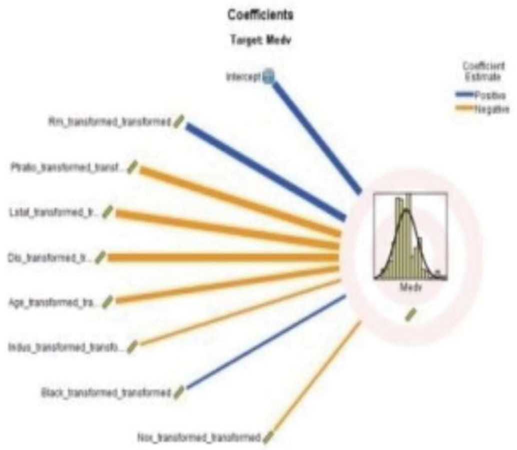


Model Summary

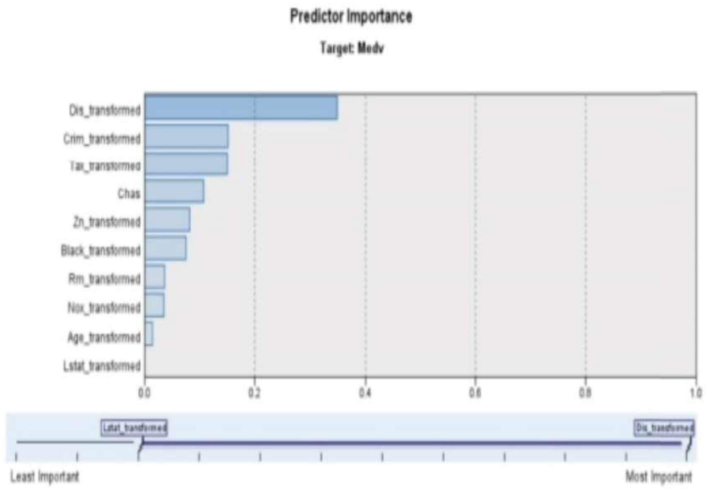
Target	Medv
Automatic Data Preparation	On
Model Selection Method	Forward Stepwise
Information Criterion	1,054.382

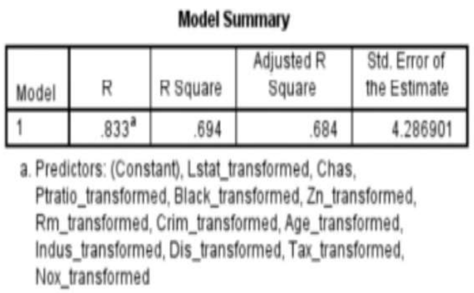
The information criterion is used to compare to models. Models with smaller information criterion values fit better.





Regression:



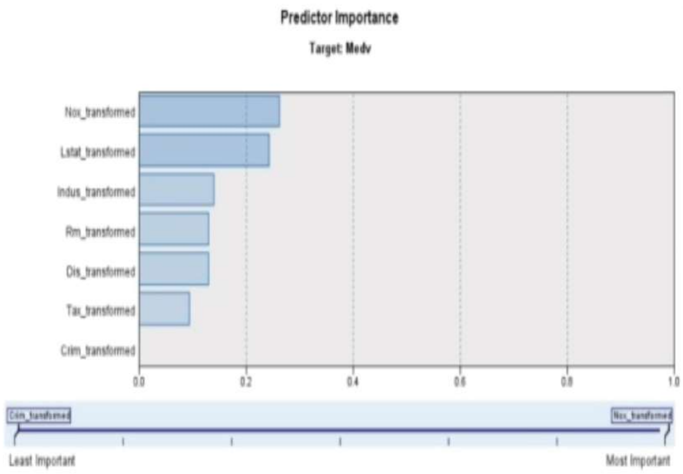


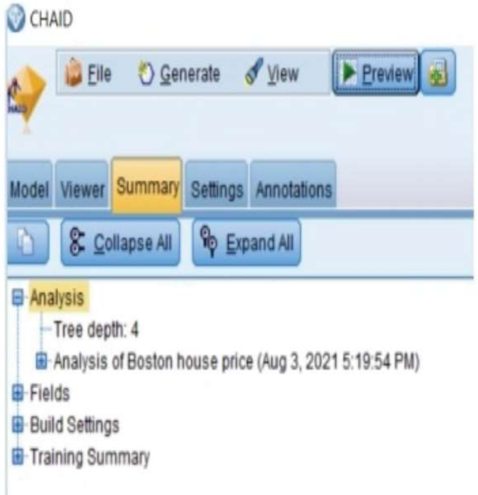
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.833[a]	.694	.684	4.286901

a. Predictors: (Constant), Lstat_transformed, Chas, Ptratio_transformed, Black_transformed, Zn_transformed, Rm_transformed, Crim_transformed, Age_transformed, Indus_transformed, Dis_transformed, Tax_transformed, Nox_transformed

CHAID:





Analysis
- Tree depth: 4
- Analysis of Boston house price (Aug 3, 2021 5:19:54 PM)

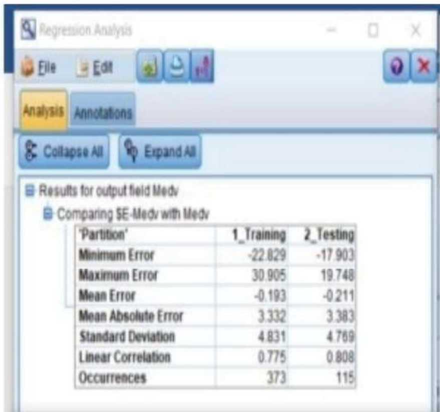
Fields

Build Settings

Training Summary

5. Evaluation

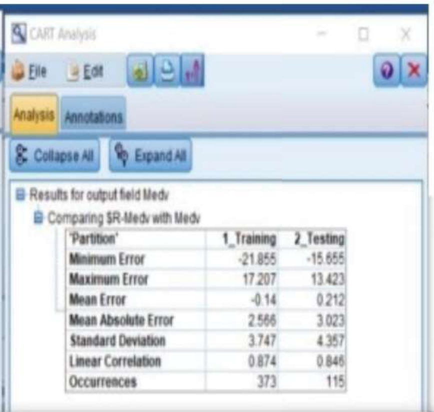
The most accurate and robust method in "One" time running is NN with 82.5% accuracy and relative Error 0.284. Other methods are less accurate or not robust (which shown with cross symbol 'X') rather than NN method. Linear Regression shows a robustness and accuracy of about 70%.



Results for output field Medv

Comparing $E-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-22.829	-17.903
Maximum Error	30.905	19.748
Mean Error	-0.193	-0.211
Mean Absolute Error	3.332	3.383
Standard Deviation	4.831	4.769
Linear Correlation	0.775	0.808
Occurrences	373	115

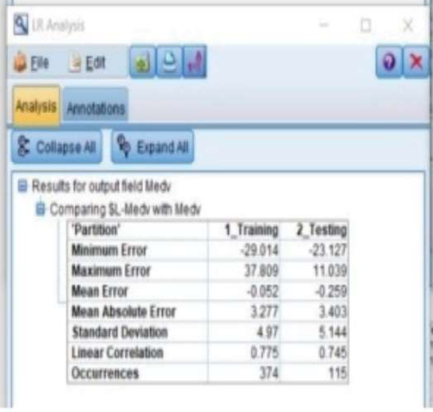


Results for output field Medv

Comparing $R-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-21.855	-15.655
Maximum Error	17.207	13.423
Mean Error	-0.14	0.212
Mean Absolute Error	2.566	3.023
Standard Deviation	3.747	4.357
Linear Correlation	0.874	0.846
Occurrences	373	115

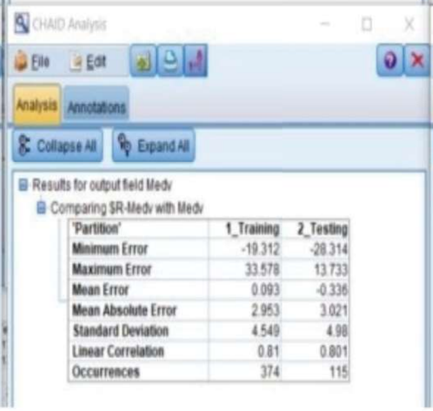
Results of Modeling on Cleaned Datasets



Results for output field Medv

Comparing $L-Medv with Medv

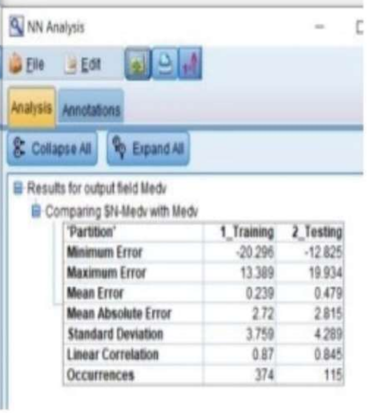
'Partition'	1_Training	2_Testing
Minimum Error	-29.014	-23.127
Maximum Error	37.809	11.039
Mean Error	-0.052	-0.259
Mean Absolute Error	3.277	3.403
Standard Deviation	4.97	5.144
Linear Correlation	0.775	0.745
Occurrences	374	115



Results for output field Medv

Comparing $R-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-19.312	-28.314
Maximum Error	33.578	13.733
Mean Error	0.093	-0.336
Mean Absolute Error	2.953	3.021
Standard Deviation	4.549	4.98
Linear Correlation	0.81	0.801
Occurrences	374	115



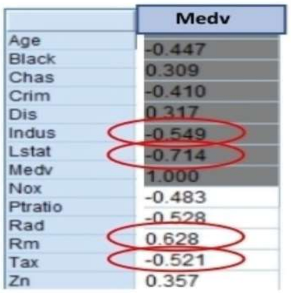
Results for output field Medv

Comparing $N-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-20.296	-12.825
Maximum Error	13.389	19.934
Mean Error	0.239	0.479
Mean Absolute Error	2.72	2.815
Standard Deviation	3.759	4.289
Linear Correlation	0.87	0.845
Occurrences	374	115

Four- and Five-Feature Extraction based on four important features seen in Model Results in Descending Order, respectively: (Dis, LSTAT, RM, Crim) and (Dis, LSTAT, RM, Crim, Ptratio)

IMPORTANCE 1:Highest 4: Least	Crim	ZN	INDUS	CHAS	NOX	RM	Age	Dis	Rad	Tax	PTRATIO	Black	LSTAT
Neural Network	3					1		4					2
CART						1	3	4					2
Linear Regression						1		4			2		3
Regression	2			4				1		3			

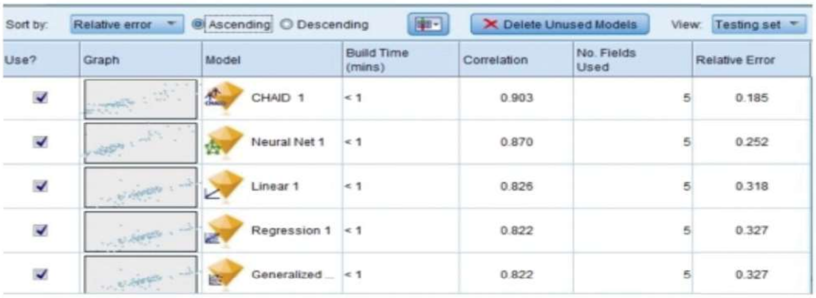


	Medv
Age	-0.447
Black	0.309
Chas	-0.410
Crim	0.317
Dis	-0.549
Indus	-0.714
Lstat	1.000
Medv	-0.483
Nox	-0.528
Ptratio	0.628
Rad	-0.521
Rm	0.357
Tax	
Zn	

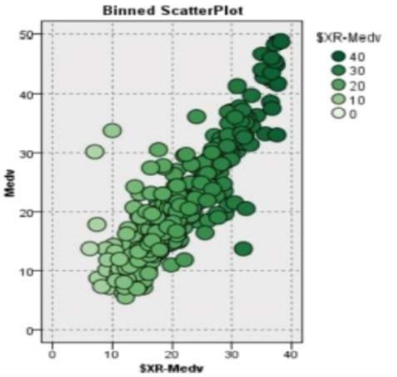
Four- and Five-Feature Extraction based on Correlation Matrix in Descending Order, respectively:

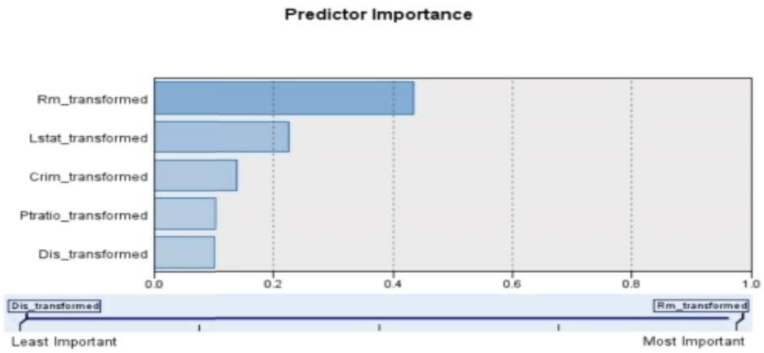
(LSTAT, RM, Indus, Ptratio) and (LSTAT, RM, Indus, Ptratio, Tax)

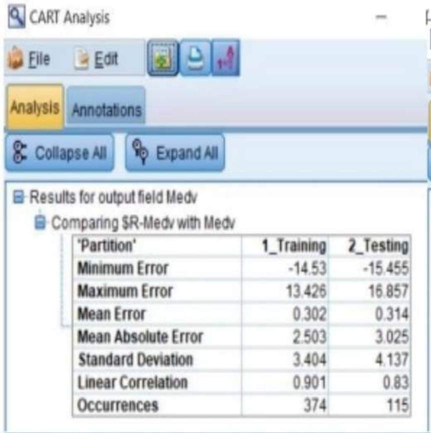
Building Models with 5 selected Features:



Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		CHAID 1	< 1	0.903	5	0.185
☑		Neural Net 1	< 1	0.870	5	0.252
☑		Linear 1	< 1	0.826	5	0.318
☑		Regression 1	< 1	0.822	5	0.327
☑		Generalized ...	< 1	0.822	5	0.327



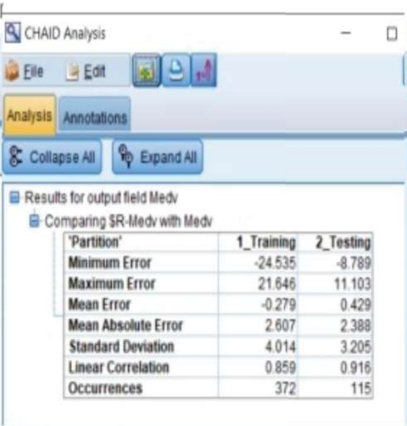




Results for output field Medv

Comparing $R-Medv with Medv

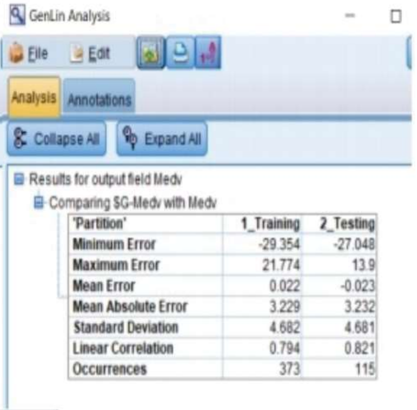
'Partition'	1_Training	2_Testing
Minimum Error	-14.53	-15.455
Maximum Error	13.426	16.857
Mean Error	0.302	0.314
Mean Absolute Error	2.503	3.025
Standard Deviation	3.404	4.137
Linear Correlation	0.901	0.83
Occurrences	374	115



Results for output field Medv

Comparing $R-Medv with Medv

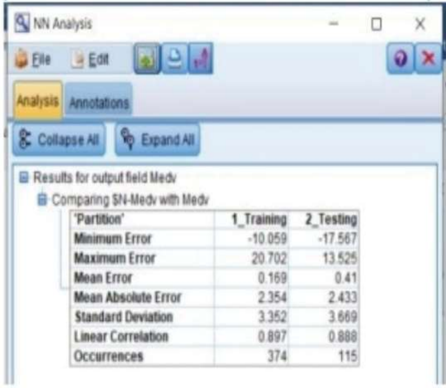
'Partition'	1_Training	2_Testing
Minimum Error	-24.535	-8.789
Maximum Error	21.646	11.103
Mean Error	-0.279	0.429
Mean Absolute Error	2.607	2.388
Standard Deviation	4.014	3.205
Linear Correlation	0.859	0.916
Occurrences	372	115



Results for output field Medv

Comparing $G-Medv with Medv

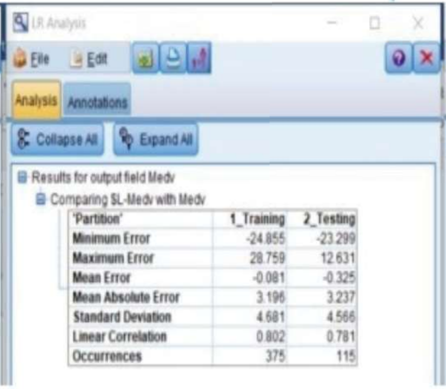
'Partition'	1_Training	2_Testing
Minimum Error	-29.354	-27.048
Maximum Error	21.774	13.9
Mean Error	0.022	-0.023
Mean Absolute Error	3.229	3.232
Standard Deviation	4.682	4.681
Linear Correlation	0.794	0.821
Occurrences	373	115



Results for output field Medv

Comparing $N-Medv with Medv

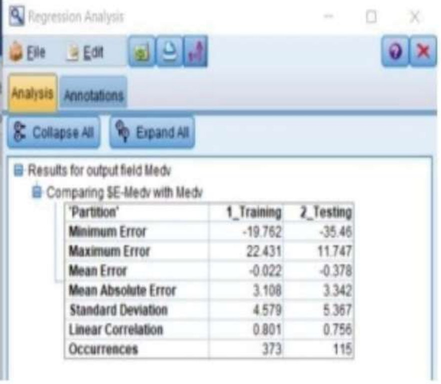
'Partition'	1_Training	2_Testing
Minimum Error	-10.059	-17.567
Maximum Error	20.702	13.525
Mean Error	0.169	0.41
Mean Absolute Error	2.354	2.433
Standard Deviation	3.352	3.669
Linear Correlation	0.897	0.888
Occurrences	374	115



Results for output field Medv

Comparing $L-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-24.855	-23.299
Maximum Error	28.759	12.631
Mean Error	-0.081	-0.325
Mean Absolute Error	3.196	3.237
Standard Deviation	4.681	4.566
Linear Correlation	0.802	0.781
Occurrences	375	115

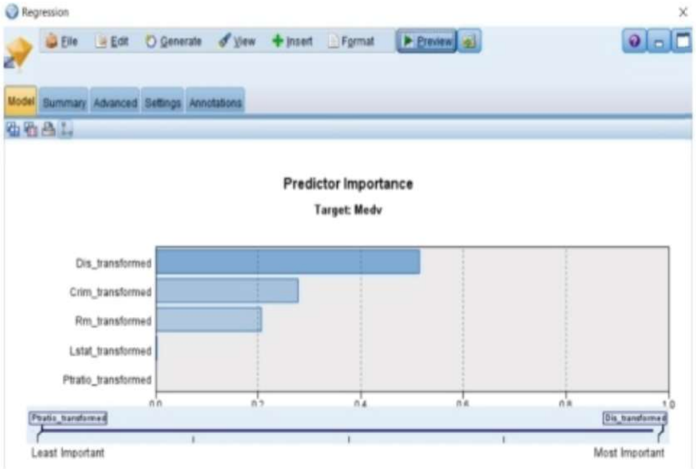


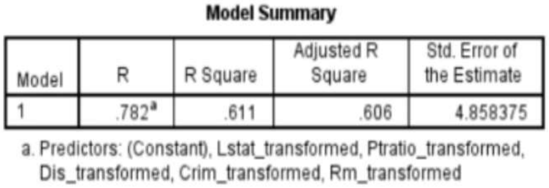
Results for output field Medv

Comparing $E-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-19.762	-35.46
Maximum Error	22.431	11.747
Mean Error	-0.022	-0.378
Mean Absolute Error	3.108	3.342
Standard Deviation	4.579	5.367
Linear Correlation	0.801	0.756
Occurrences	373	115

Regression:



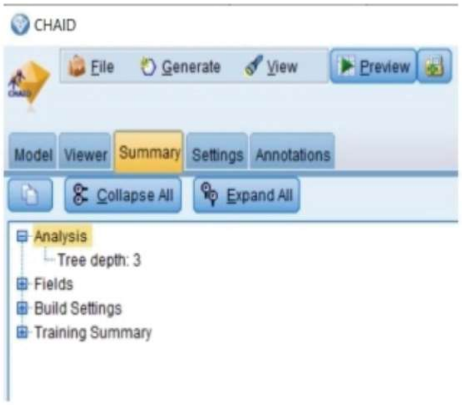


Model Summary

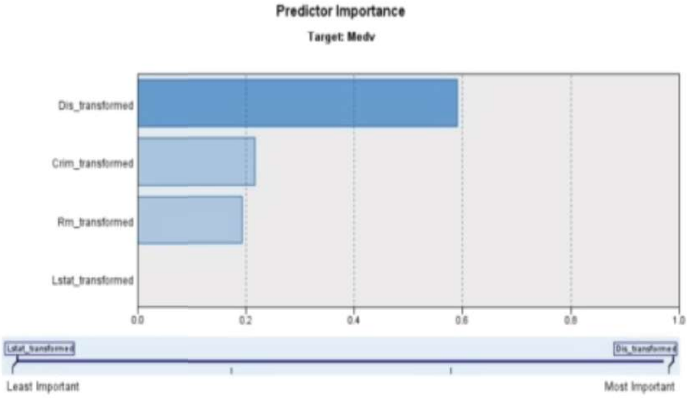
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782[a]	.611	.606	4.858375

a. Predictors: (Constant), Lstat_transformed, Ptratio_transformed, Dis_transformed, Crim_transformed, Rm_transformed

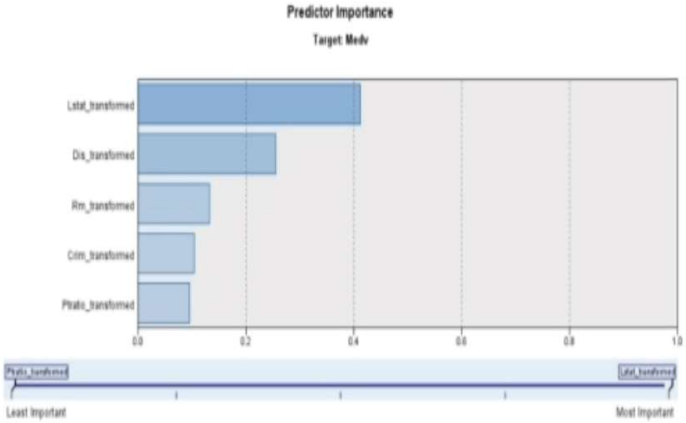
CHAID:

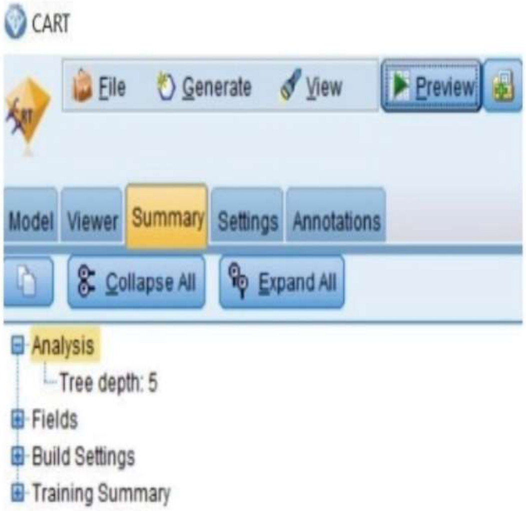


Analysis
Tree depth: 3
Fields
Build Settings
Training Summary



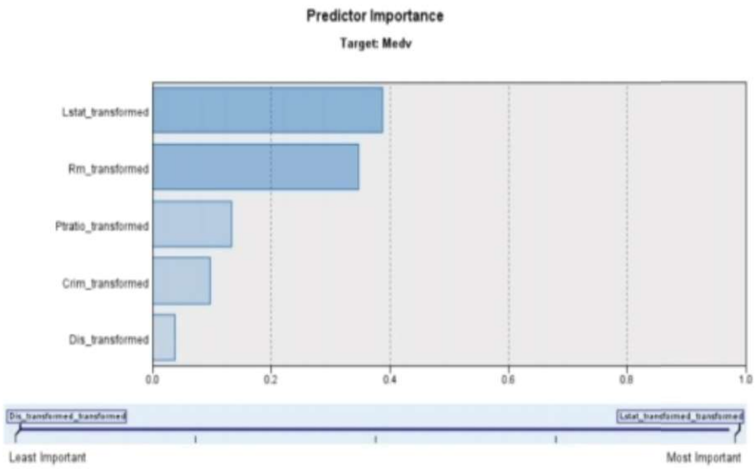
CART:

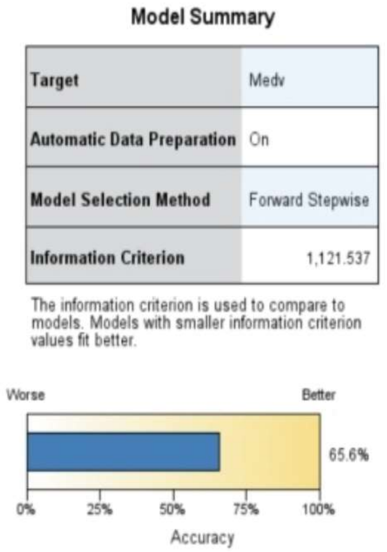




Analysis
Tree depth: 5
Fields
Build Settings
Training Summary

Linear Regression:



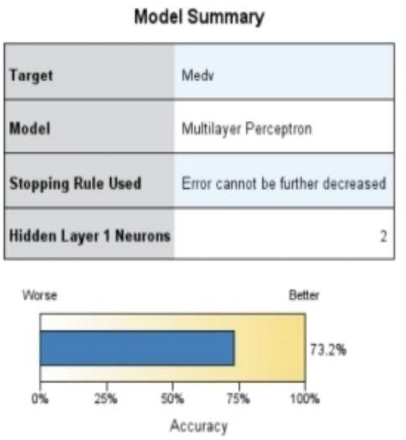


Model Summary

Target	Medv
Automatic Data Preparation	On
Model Selection Method	Forward Stepwise
Information Criterion	1,121.537

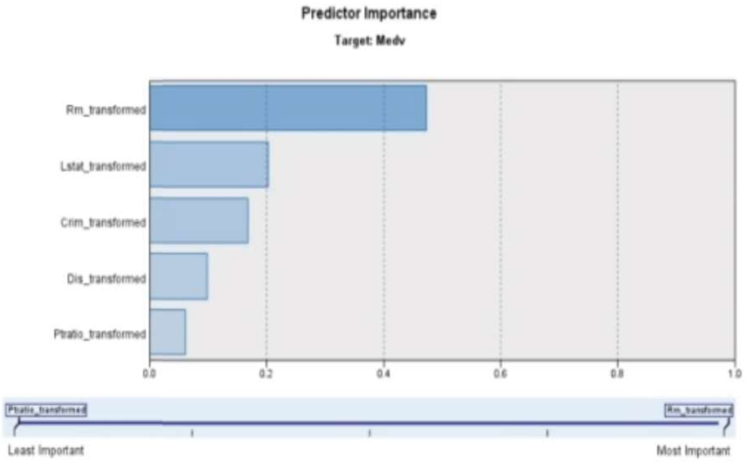
The information criterion is used to compare to models. Models with smaller information criterion values fit better.

Neural Network:

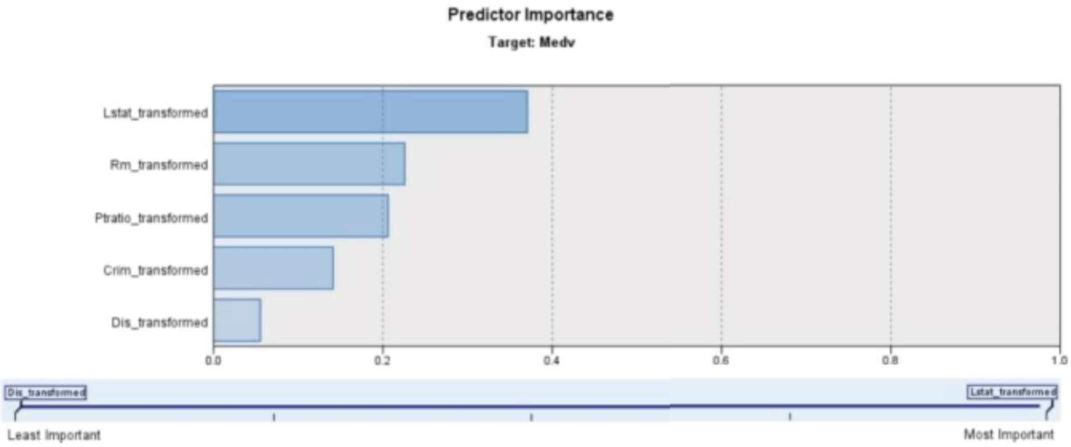


Model Summary

Target	Medv
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	2

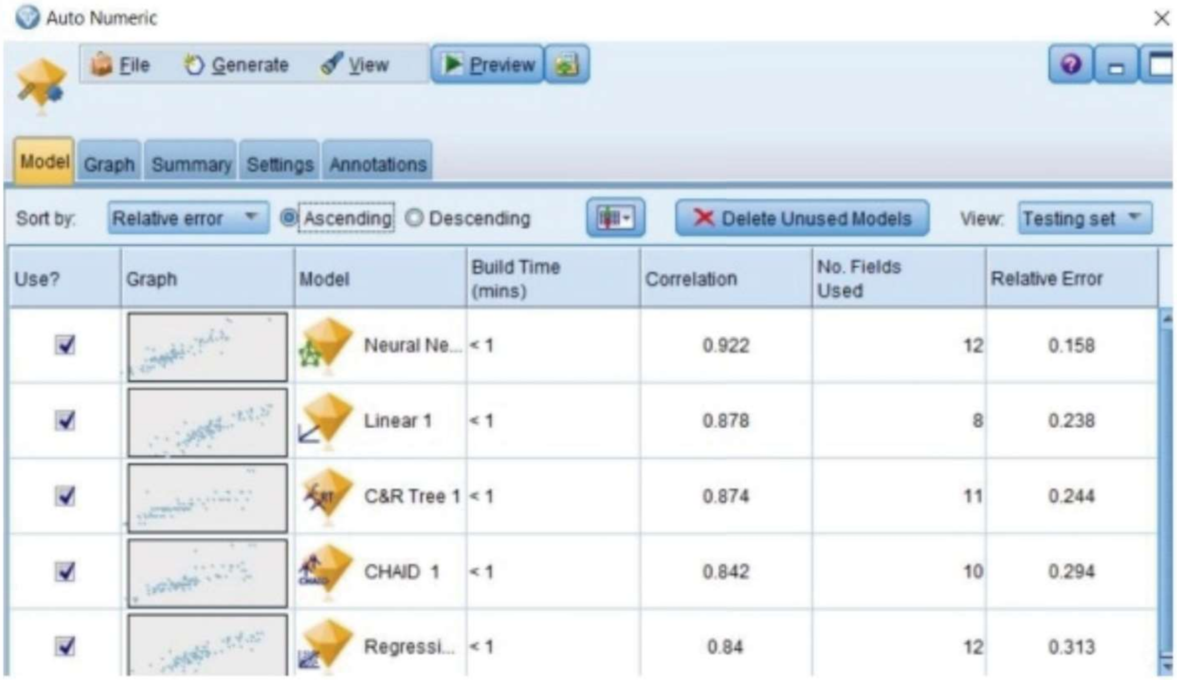


Generlized Linear Model:



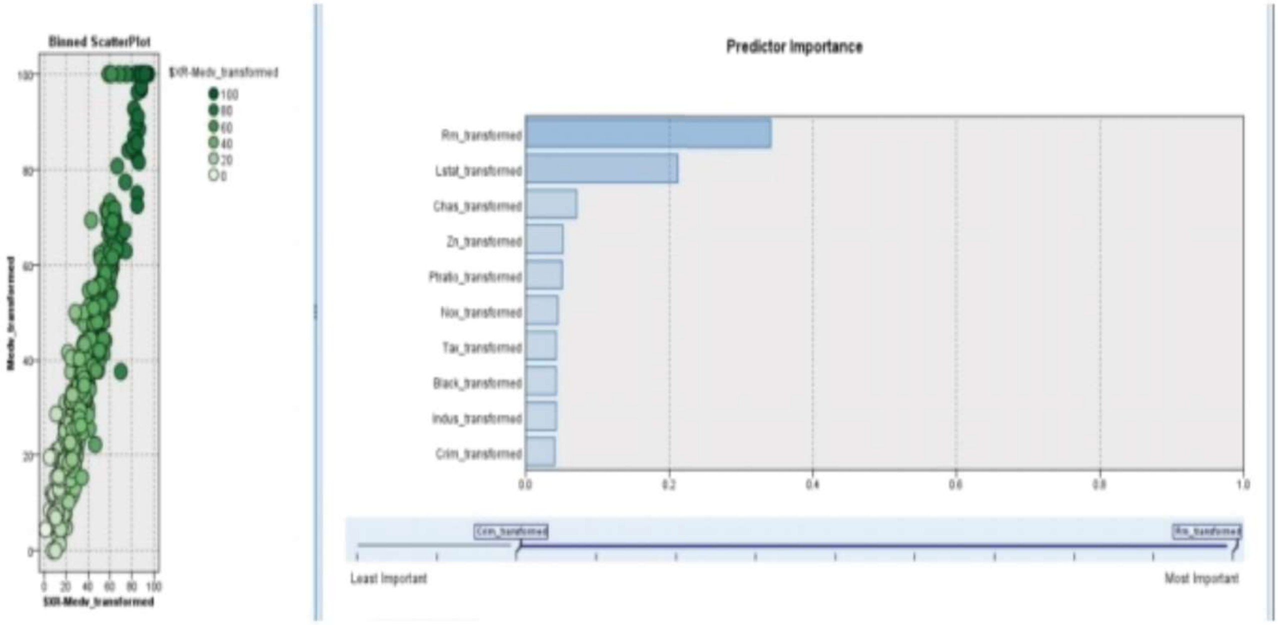
حالا ساخت مدل رو بدون clean کردن دیتا در نظر می‌گیریم و دقت ها رو می سنجیم

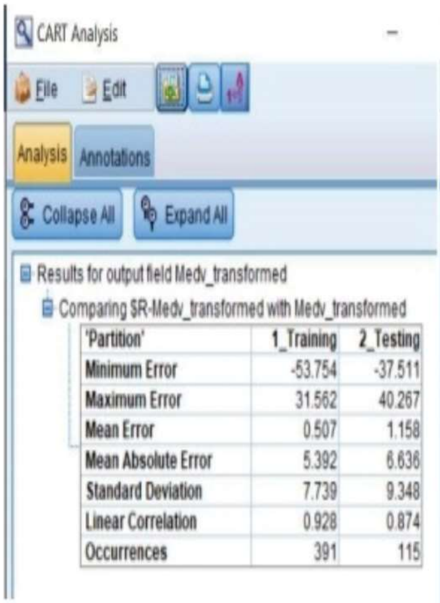
Building Models on Uncleaned Dataset:



Sort by: Relative error ● Ascending ○ Descending View: Testing set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Neural Ne...	< 1	0.922	12	0.158
☑		Linear 1	< 1	0.878	8	0.238
☑		C&R Tree 1	< 1	0.874	11	0.244
☑		CHAID 1	< 1	0.842	10	0.294
☑		Regressi...	< 1	0.84	12	0.313

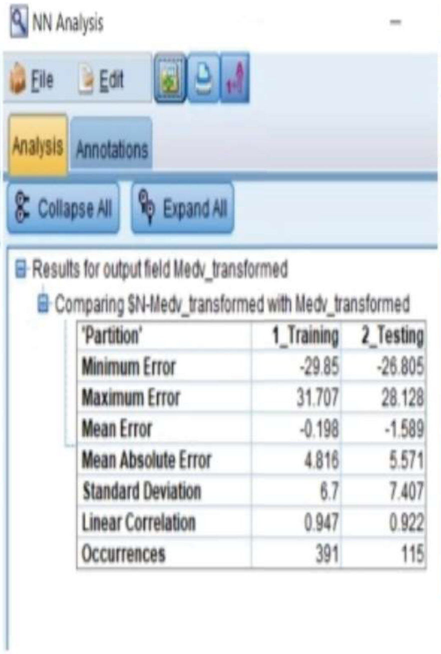




Results for output field Medv_transformed

Comparing $R-Medv_transformed with Medv_transformed

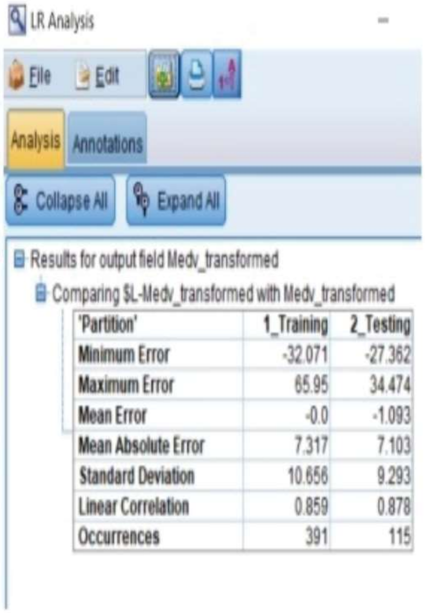
'Partition'	1_Training	2_Testing
Minimum Error	-53.754	-37.511
Maximum Error	31.562	40.267
Mean Error	0.507	1.158
Mean Absolute Error	5.392	6.636
Standard Deviation	7.739	9.348
Linear Correlation	0.928	0.874
Occurrences	391	115



Results for output field Medv_transformed

Comparing $N-Medv_transformed with Medv_transformed

'Partition'	1_Training	2_Testing
Minimum Error	-29.85	-26.805
Maximum Error	31.707	28.128
Mean Error	-0.198	-1.589
Mean Absolute Error	4.816	5.571
Standard Deviation	6.7	7.407
Linear Correlation	0.947	0.922
Occurrences	391	115



Results for output field Medv_transformed

Comparing $L-Medv_transformed with Medv_transformed

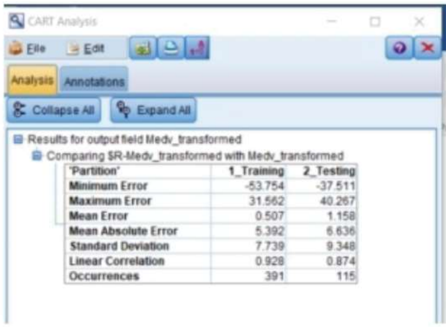
'Partition'	1_Training	2_Testing
Minimum Error	-32.071	-27.362
Maximum Error	65.95	34.474
Mean Error	-0.0	-1.093
Mean Absolute Error	7.317	7.103
Standard Deviation	10.656	9.293
Linear Correlation	0.859	0.878
Occurrences	391	115

- **Comparison:**

Four Top Important Features

- Uncleaned Data: RM, LSTAT, Chas, ZN
- Cleaned Data:
  - All Features: RM, LSTAT, Indus, Ptratio
  - Less Features: RM, LSTAT, Crim, Ptratio

- Model Analysis on Uncleaned Data

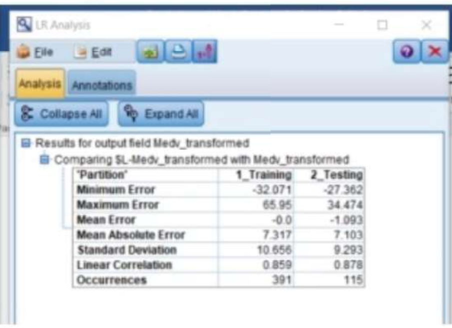


Results for output field Medv_transformed

Comparing $R-Medv_transformed with Medv_transformed

'Partition'	1_Training	2_Testing
Minimum Error	-53.754	-37.511
Maximum Error	31.562	40.267
Mean Error	0.507	1.158
Mean Absolute Error	5.392	6.636
Standard Deviation	7.739	9.348
Linear Correlation	0.928	0.874
Occurrences	391	115

X

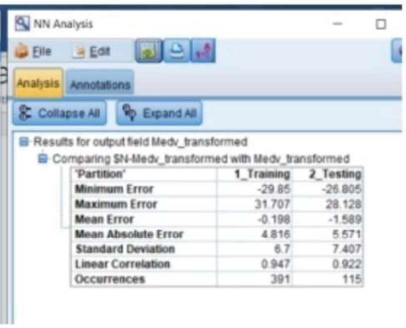


Results for output field Medv_transformed

Comparing $L-Medv_transformed with Medv_transformed

'Partition'	1_Training	2_Testing
Minimum Error	-32.071	-27.362
Maximum Error	65.95	34.474
Mean Error	-0.0	-1.093
Mean Absolute Error	7.317	7.103
Standard Deviation	10.656	9.293
Linear Correlation	0.859	0.878
Occurrences	391	115

X



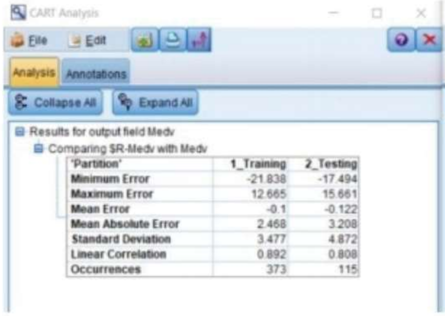
Results for output field Medv_transformed

Comparing $N-Medv_transformed with Medv_transformed

'Partition'	1_Training	2_Testing
Minimum Error	-29.85	-26.805
Maximum Error	31.707	28.128
Mean Error	-0.198	-1.589
Mean Absolute Error	4.816	5.571
Standard Deviation	6.7	7.407
Linear Correlation	0.947	0.922
Occurrences	391	115

X

- Model Analysis on Less Features Datasets

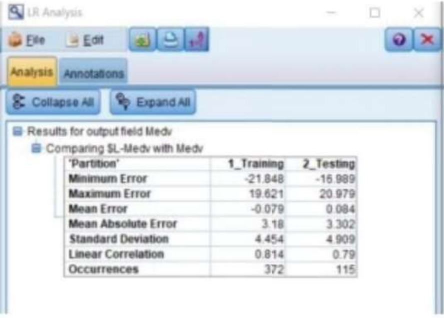


Results for output field Medv

Comparing $R-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-21.838	-17.494
Maximum Error	12.665	15.661
Mean Error	-0.1	-0.122
Mean Absolute Error	2.468	3.208
Standard Deviation	3.477	4.872
Linear Correlation	0.892	0.808
Occurrences	373	115

√

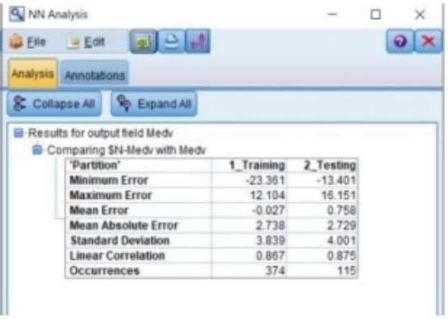


Results for output field Medv

Comparing $L-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-21.848	-16.989
Maximum Error	19.621	20.979
Mean Error	-0.079	0.084
Mean Absolute Error	3.18	3.302
Standard Deviation	4.454	4.909
Linear Correlation	0.814	0.79
Occurrences	372	115

√



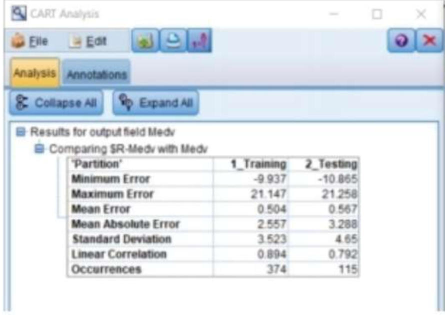
Results for output field Medv

Comparing $N-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-23.361	-13.401
Maximum Error	12.104	16.151
Mean Error	-0.027	0.758
Mean Absolute Error	2.738	2.729
Standard Deviation	3.839	4.001
Linear Correlation	0.867	0.875
Occurrences	374	115

√

- Model Analysis on Cleaned Data with all Features

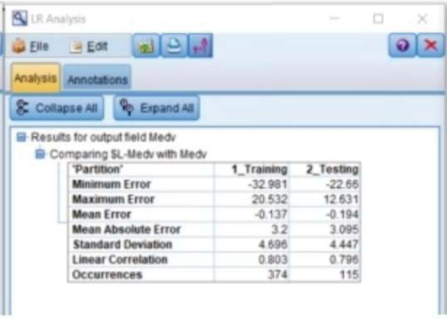


Results for output field Medv

Comparing $R-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-9.937	-10.865
Maximum Error	21.147	21.258
Mean Error	0.504	0.567
Mean Absolute Error	2.557	3.288
Standard Deviation	3.523	4.65
Linear Correlation	0.894	0.792
Occurrences	374	115

√

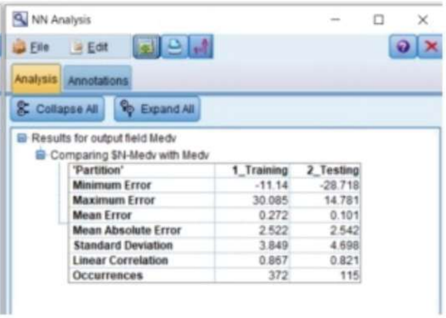


Results for output field Medv

Comparing $L-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-32.981	-22.66
Maximum Error	20.532	12.631
Mean Error	-0.137	-0.194
Mean Absolute Error	3.2	3.095
Standard Deviation	4.696	4.447
Linear Correlation	0.803	0.796
Occurrences	374	115

√



Results for output field Medv

Comparing $N-Medv with Medv

'Partition'	1_Training	2_Testing
Minimum Error	-11.14	-28.718
Maximum Error	30.085	14.781
Mean Error	0.272	0.101
Mean Absolute Error	2.522	2.542
Standard Deviation	3.849	4.698
Linear Correlation	0.867	0.821
Occurrences	372	115

√