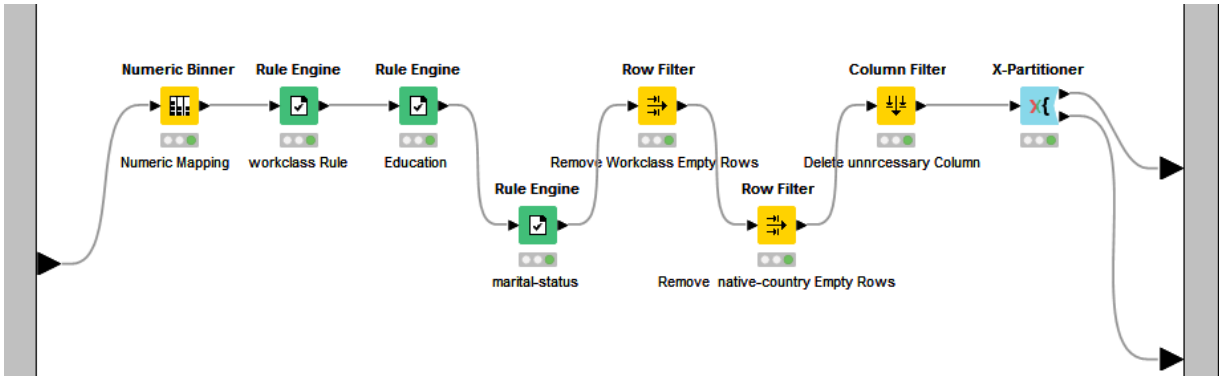
Heidar Zirak

Aden Abreham

a. Classification

Data preparation

The data prepared for association mining by removing some attributes, mapping some continuous attributes to nominal attributes, reducing number of levels of some nominal attributes, and removing empty data.



Removing some attributes:

We can remove education_num (Highest level of education in numerical form) as the data can be found from other field (for example education).and also field fnlwgt can be removed.

Mapping some attributes to another:

Divide age value to four level:

17-25 → Young Adult

26-45 → Early Middle

46-65 → Late Middle

66-90 → Late Adulthood

Divide hours per week value to four level:

.. -25 → Part Time

25-40 →Full Time

40-60 → Over Time

60-.. → Too Much

<u>The mean of the fields is battained from Statistics Node:</u>

Mean of capital_gain =1079

We divide values of capital_gain to 3 levels:

capital_gain = 0 → None

0<capital_gain<1079 → Small

1079 > capital_gain → Big

Mean of capital_loss =87

We divide values of capital_loss to 3 levels:

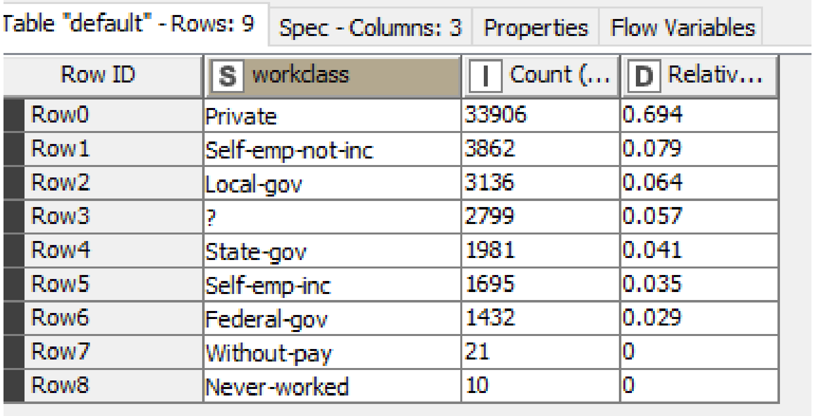
capital_loss = 0 → None

0<capital_loss <87 → Small

87 > capital_loss → Big

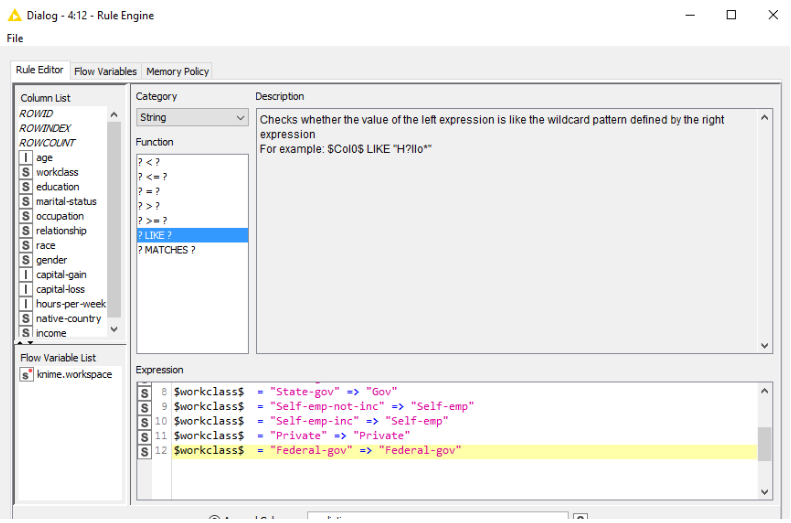
<u>Reducing number of levels of some nominal attributes:</u>

We saw that "Never-worked" and "Without-Pay" are both very small groups, and they are likely very similar, so we will combine them to form a "Not-Working" category.



Row ID	S workclass	I Count (...	D Relativ...
Row0	Private	33906	0.694
Row1	Self-emp-not-inc	3862	0.079
Row2	Local-gov	3136	0.064
Row3	?	2799	0.057
Row4	State-gov	1981	0.041
Row5	Self-emp-inc	1695	0.035
Row6	Federal-gov	1432	0.029
Row7	Without-pay	21	0
Row8	Never-worked	10	0

With the rule engine node the workClass will be reduce to shared category as below :



Category: String

Function: ? LIKE ?

Description: Checks whether the value of the left expression is like the wildcard pattern defined by the right expression
For example: $Col0$ LIKE "H?llo*"

```
$workclass$  = "State-gov" => "Gov"
$workclass$  = "Self-emp-not-inc" => "Self-emp"
$workclass$  = "Self-emp-inc" => "Self-emp"
$workclass$  = "Private" => "Private"
$workclass$  = "Federal-gov" => "Federal-gov"
```

With the rule engine node, the education will be reduce to shared category as below :

```
$education$  = "Preschool" => "GiveUp"
$education$  = "1st-4th"=> "GiveUp"
$education$  = "5th-6th"=> "GiveUp"
$education$  = "7th-8th"=> "GiveUp"
$education$  = "9th"=> "GiveUp"
$education$  = "10th"=> "GiveUp"
$education$  = "11th"=> "GiveUp"
$education$  = "12th"=> "GiveUp"
$education$  = "Assoc-acdm"=> "Associates"
$education$  = "Assoc-voc"=> "Associates"
$education$  = "HS-grad"=> "HighSchoolGraduate"
$education$  = "Some-college"=> "HighSchoolGraduate"
$education$  = "Prof-school"=> "Prof-school"
$education$  = "Bachelors"=> "Bachelors"
$education$  = "Masters"=> "Masters"
$education$  = "Doctorate"=> "Doctorate"
```

marital-status

```
$marital-status$= "Married-AF-spouse"=> "Married"
$marital-status$= "Married-civ-spouse"=> "Married"
$marital-status$= "Married-spouse-absent"=> "Not-married"
$marital-status$= "Separated"=> "Not-married"
$marital-status$= "Divorced"=> "Not-married"
$marital-status$= "Widowed"=> "Widowed"
$marital-status$= "Never-married"=> "Never-married"
```

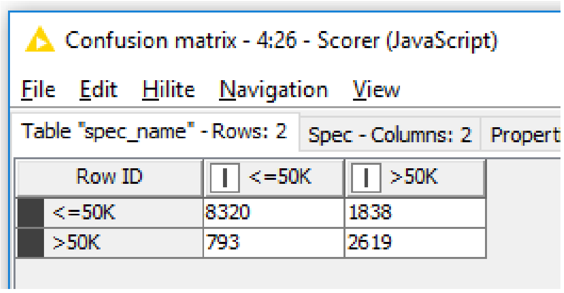
Remove empty data

Those of record in "Workclass"And "native-country " which had some missing Data, removed from Table.

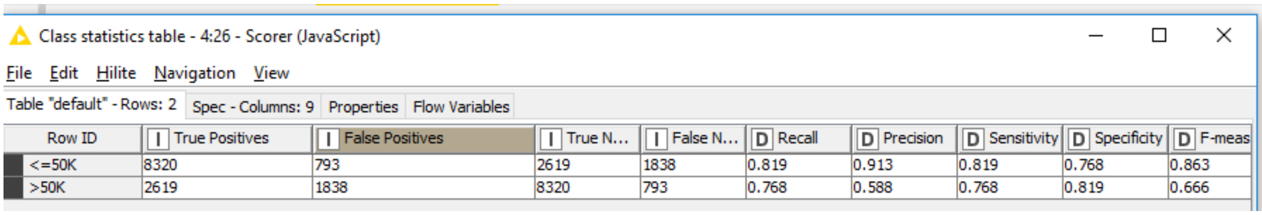
Classifier Analyze:

Three Different machine learning classifier selected as below and for each of them the information is depicted.
According to cited information the **Naive Base Classifier** with about 84% overall accuracy shows better accuracy among the other training model. Also, Confusion matrix depict that the prediction income for <=50k is ,9494 true Guess and 1521 False Guess and about the >50 k ,we have 1891 and 664 for True and False Guess.

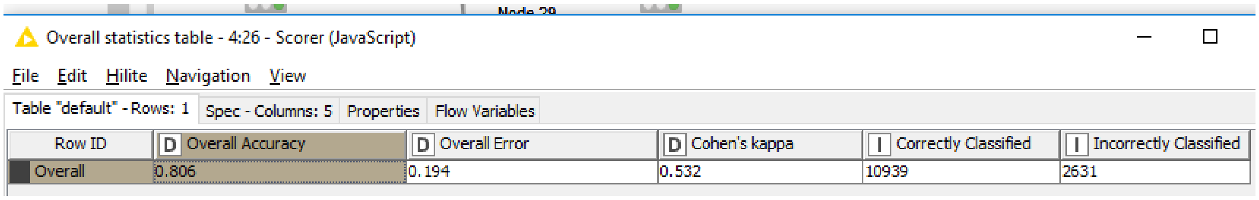
Decision Tree Classifier


Confusion matrix - 4:26 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "spec_name" - Rows: 2
Spec - Columns: 2
Propert

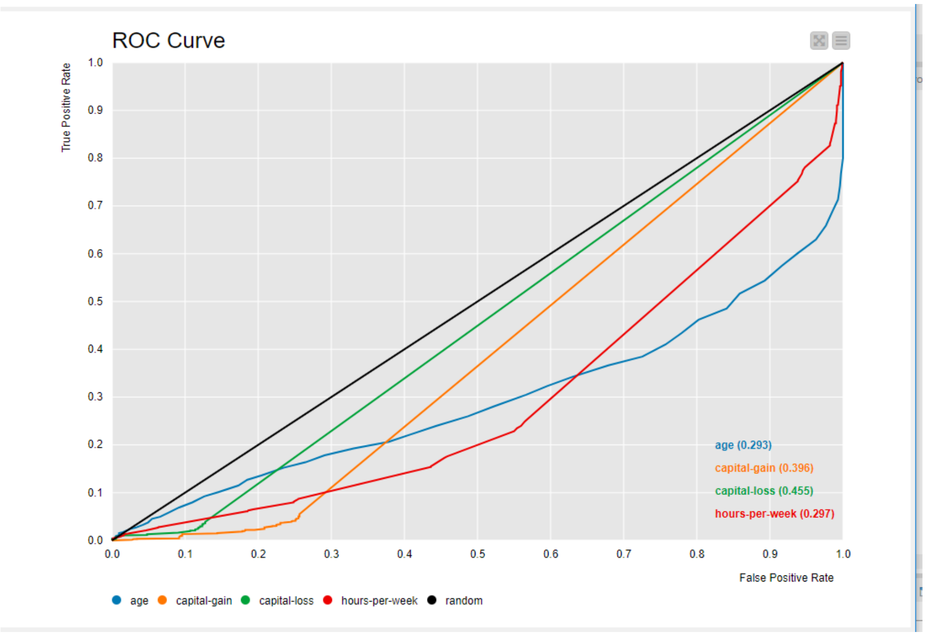
Row ID	I <=50K	I >50K
<=50K	8320	1838
>50K	793	2619


Class statistics table - 4:26 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 2
Spec - Columns: 9
Properties
Flow Variables

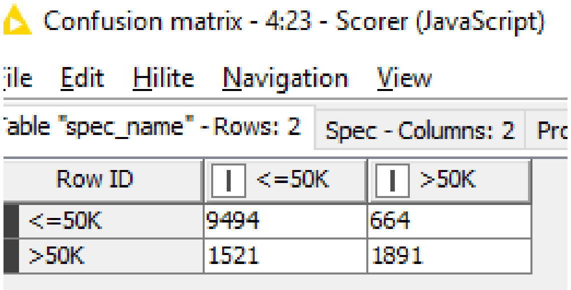
Row ID	I True Positives	I False Positives	I True N...	I False N...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas
<=50K	8320	793	2619	1838	0.819	0.913	0.819	0.768	0.863
>50K	2619	1838	8320	793	0.768	0.588	0.768	0.819	0.666


Node 29
Overall statistics table - 4:26 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 1
Spec - Columns: 5
Properties
Flow Variables

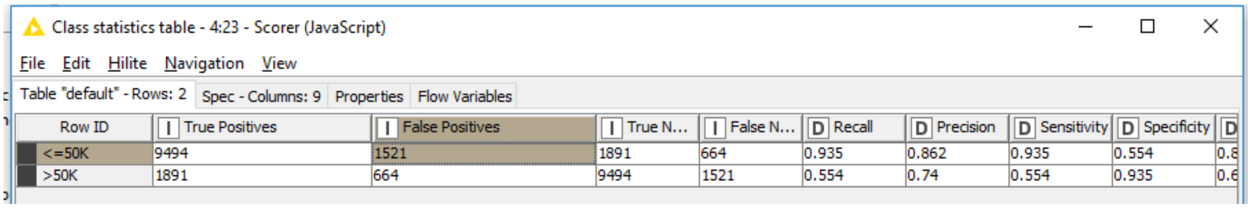
Row ID	D Overall Accuracy	D Overall Error	D Cohen's kappa	I Correctly Classified	I Incorrectly Classified
Overall	0.806	0.194	0.532	10939	2631



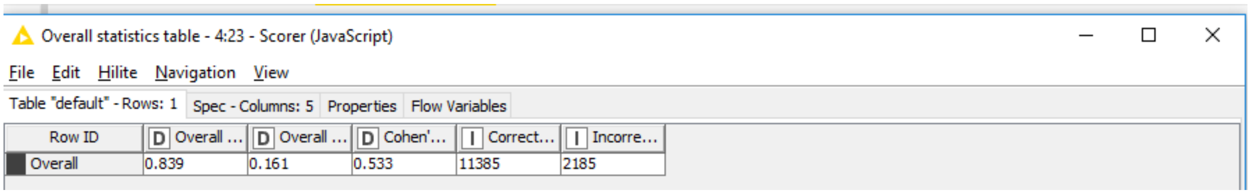
Naive Base Classifier


Confusion matrix - 4:23 - Scorer (JavaScript)
ile Edit Hilite Navigation View
able "spec_name" - Rows: 2
Spec - Columns: 2
Pro

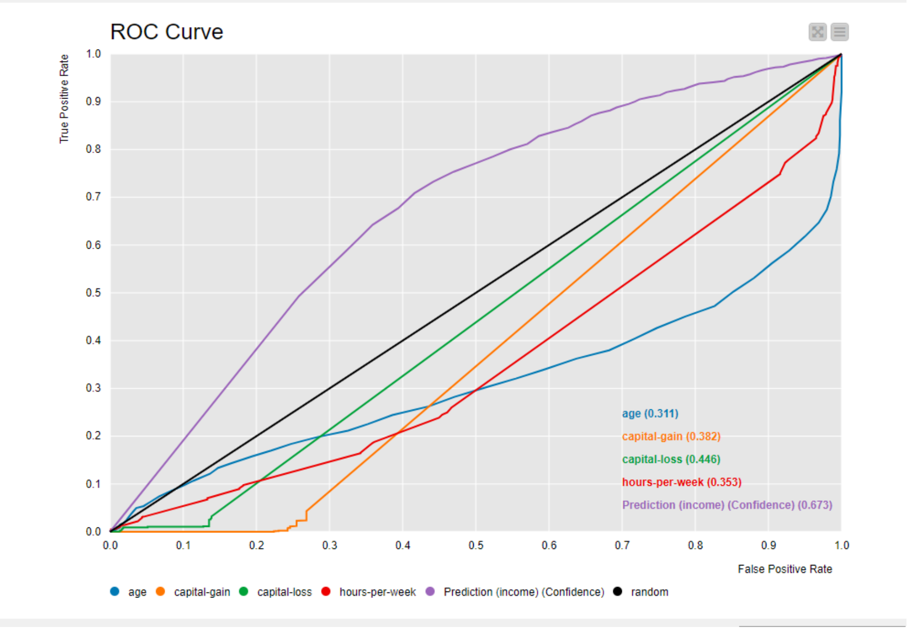
Row ID	I <=50K	I >50K
<=50K	9494	664
>50K	1521	1891


Class statistics table - 4:23 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 2
Spec - Columns: 9
Properties
Flow Variables

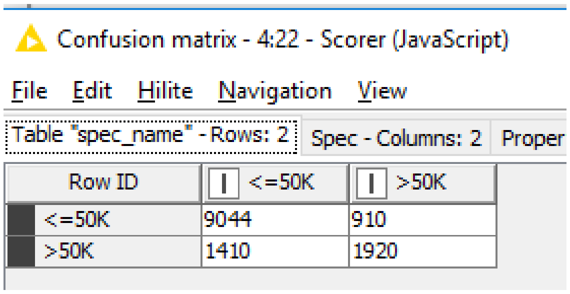
Row ID	I True Positives	I False Positives	I True N...	I False N...	D Recall	D Precision	D Sensitivity	D Specificity	D
<=50K	9494	1521	1891	664	0.935	0.862	0.935	0.554	0.8
>50K	1891	664	9494	1521	0.554	0.74	0.554	0.935	0.6


Overall statistics table - 4:23 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 1
Spec - Columns: 5
Properties
Flow Variables

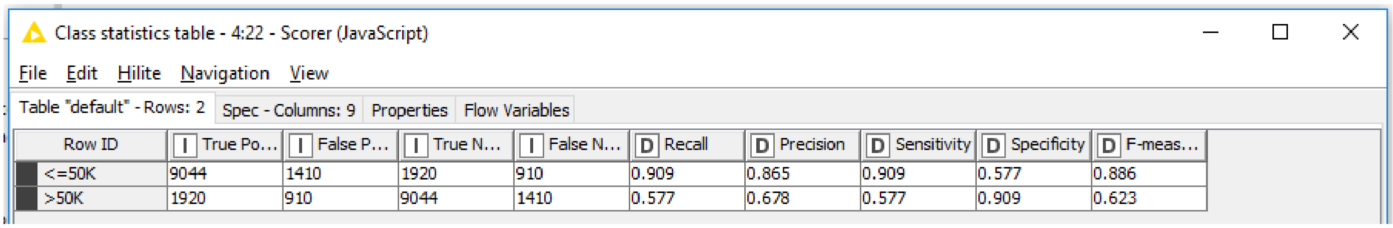
Row ID	D Overall ...	D Overall ...	D Cohen'...	I Correct...	I Incorre...
Overall	0.839	0.161	0.533	11385	2185



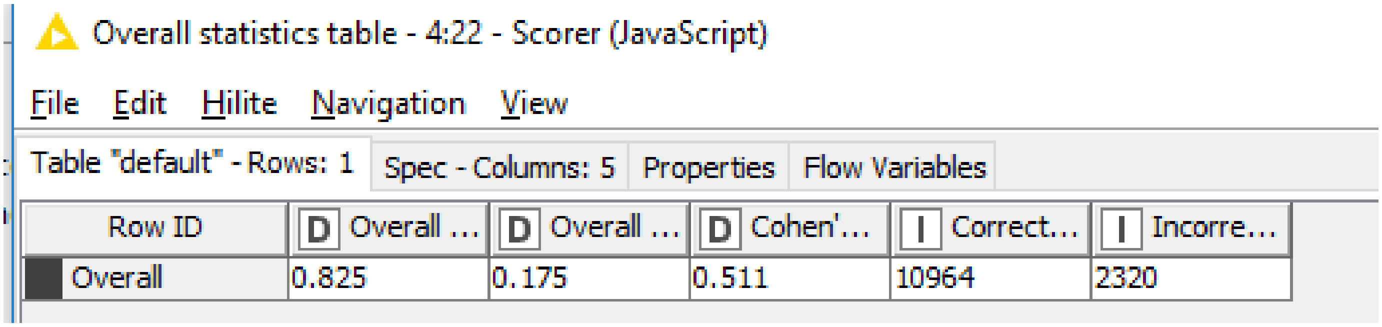
Random Forest Classifier


Confusion matrix - 4:22 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "spec_name" - Rows: 2
Spec - Columns: 2
Proper

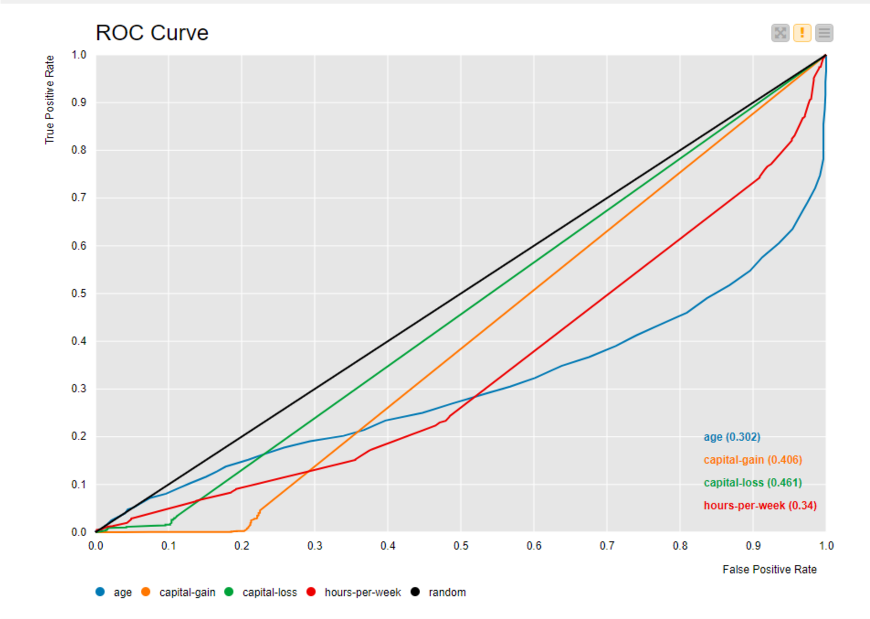
Row ID	I <=50K	I >50K
<=50K	9044	910
>50K	1410	1920


Class statistics table - 4:22 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 2
Spec - Columns: 9
Properties
Flow Variables

Row ID	I True Po…	I False P…	I True N…	I False N…	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas…
<=50K	9044	1410	1920	910	0.909	0.865	0.909	0.577	0.886
>50K	1920	910	9044	1410	0.577	0.678	0.577	0.909	0.623

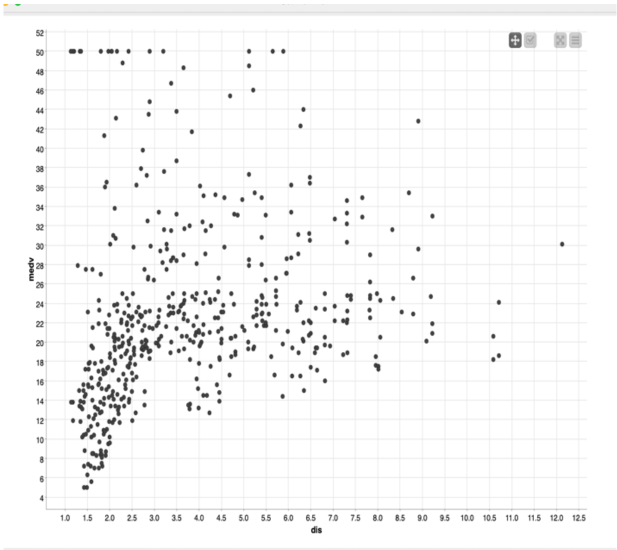

Overall statistics table - 4:22 - Scorer (JavaScript)
File Edit Hilite Navigation View
Table "default" - Rows: 1
Spec - Columns: 5
Properties
Flow Variables

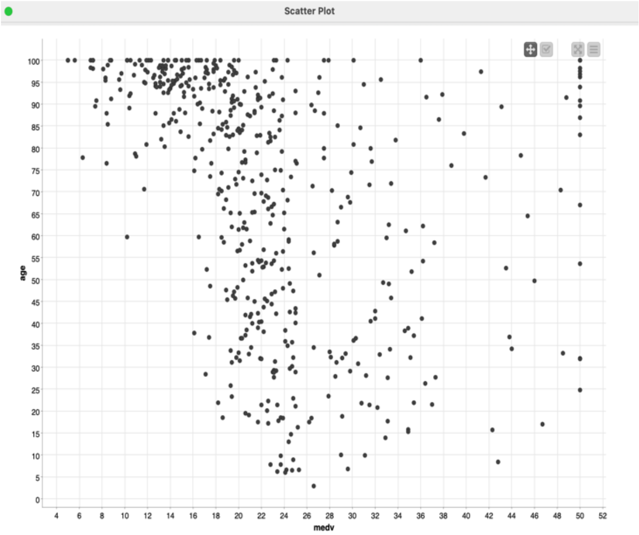
Row ID	D Overall …	D Overall …	D Cohen'…	I Correct…	I Incorre…
Overall	0.825	0.175	0.511	10964	2320

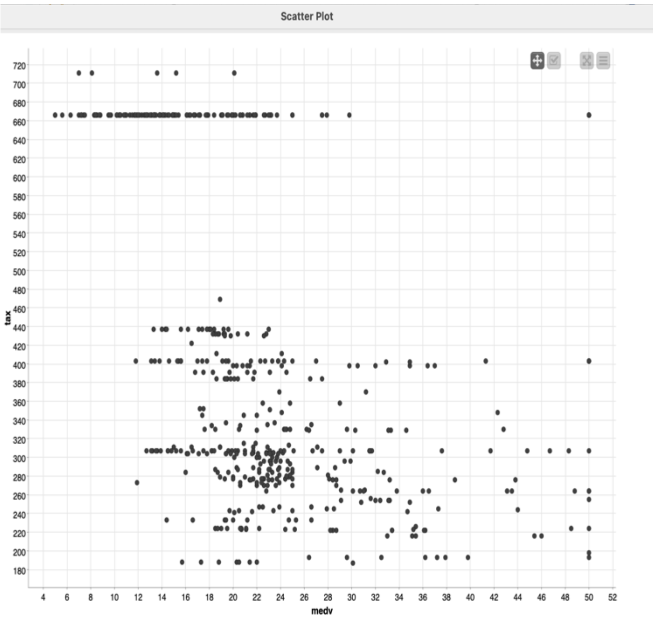


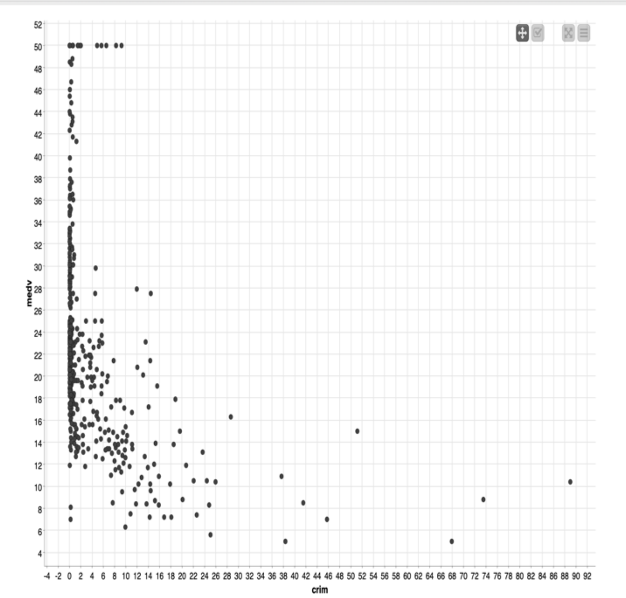
a. Regression

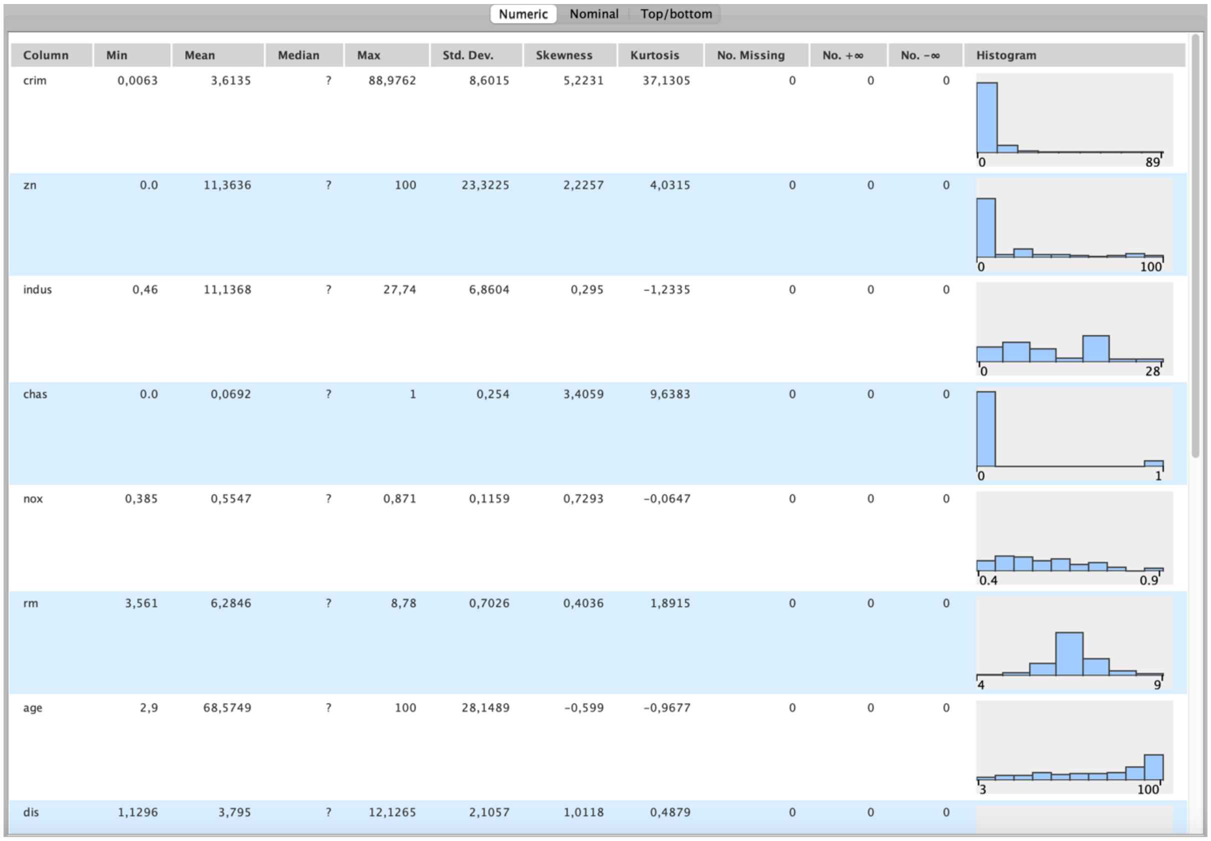
Some of the attribute's relation with the target attribute MEDV in visual illustration.



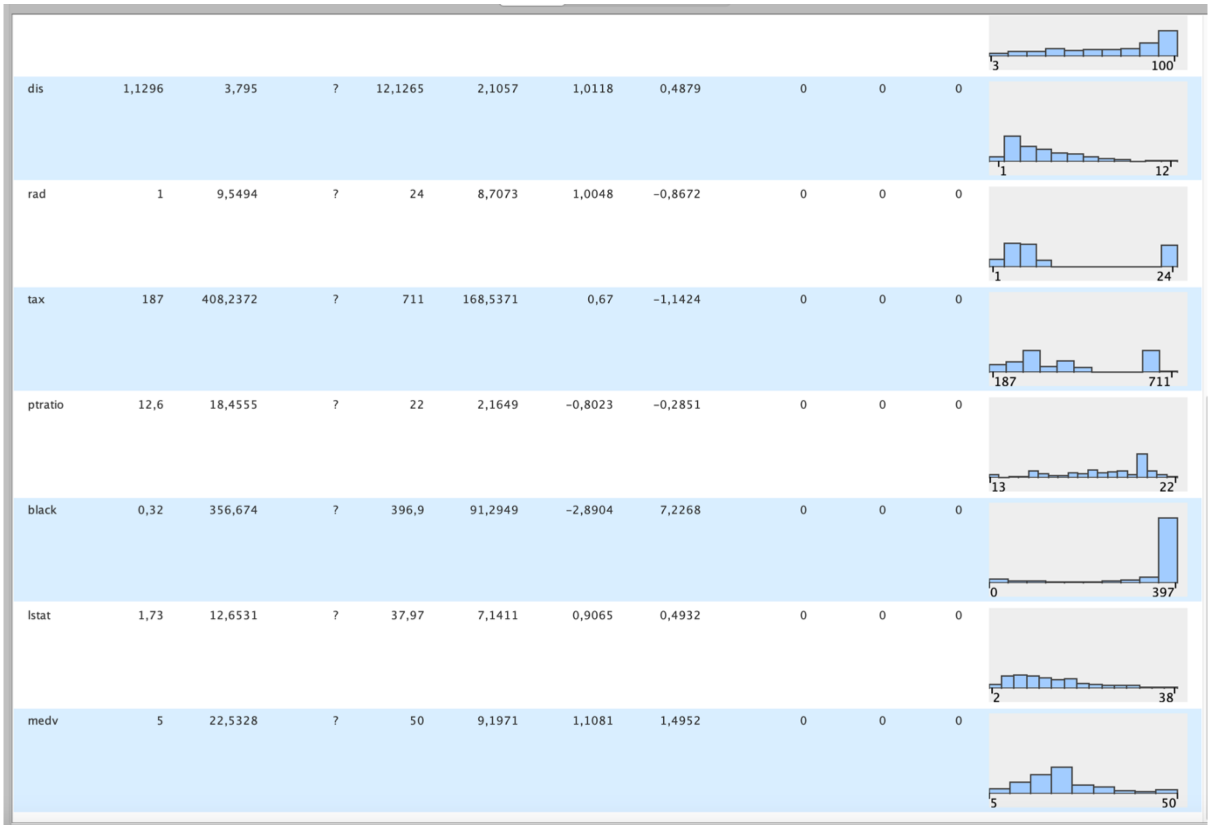








Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. −∞	Histogram
crim	0,0063	3,6135	?	88,9762	8,6015	5,2231	37,1305	0	0	0	0 89
zn	0.0	11,3636	?	100	23,3225	2,2257	4,0315	0	0	0	0 100
indus	0,46	11,1368	?	27,74	6,8604	0,295	−1,2335	0	0	0	0 28
chas	0.0	0,0692	?	1	0,254	3,4059	9,6383	0	0	0	0 1
nox	0,385	0,5547	?	0,871	0,1159	0,7293	−0,0647	0	0	0	0.4 0.9
rm	3,561	6,2846	?	8,78	0,7026	0,4036	1,8915	0	0	0	4 9
age	2,9	68,5749	?	100	28,1489	−0,599	−0,9677	0	0	0	3 100
dis	1,1296	3,795	?	12,1265	2,1057	1,0118	0,4879	0	0	0	



dis	1,1296	3,795	?	12,1265	2,1057	1,0118	0,4879	0	0	0
rad	1	9,5494	?	24	8,7073	1,0048	-0,8672	0	0	0
tax	187	408,2372	?	711	168,5371	0,67	-1,1424	0	0	0
ptratio	12,6	18,4555	?	22	2,1649	-0,8023	-0,2851	0	0	0
black	0,32	356,674	?	396,9	91,2949	-2,8904	7,2268	0	0	0
lstat	1,73	12,6531	?	37,97	7,1411	0,9065	0,4932	0	0	0
medv	5	22,5328	?	50	9,1971	1,1081	1,4952	0	0	0

Preparing the data for modelling

The Boston Housing Market dataset has many features so, the dimensional reduction is measure based.

Ration of missing value

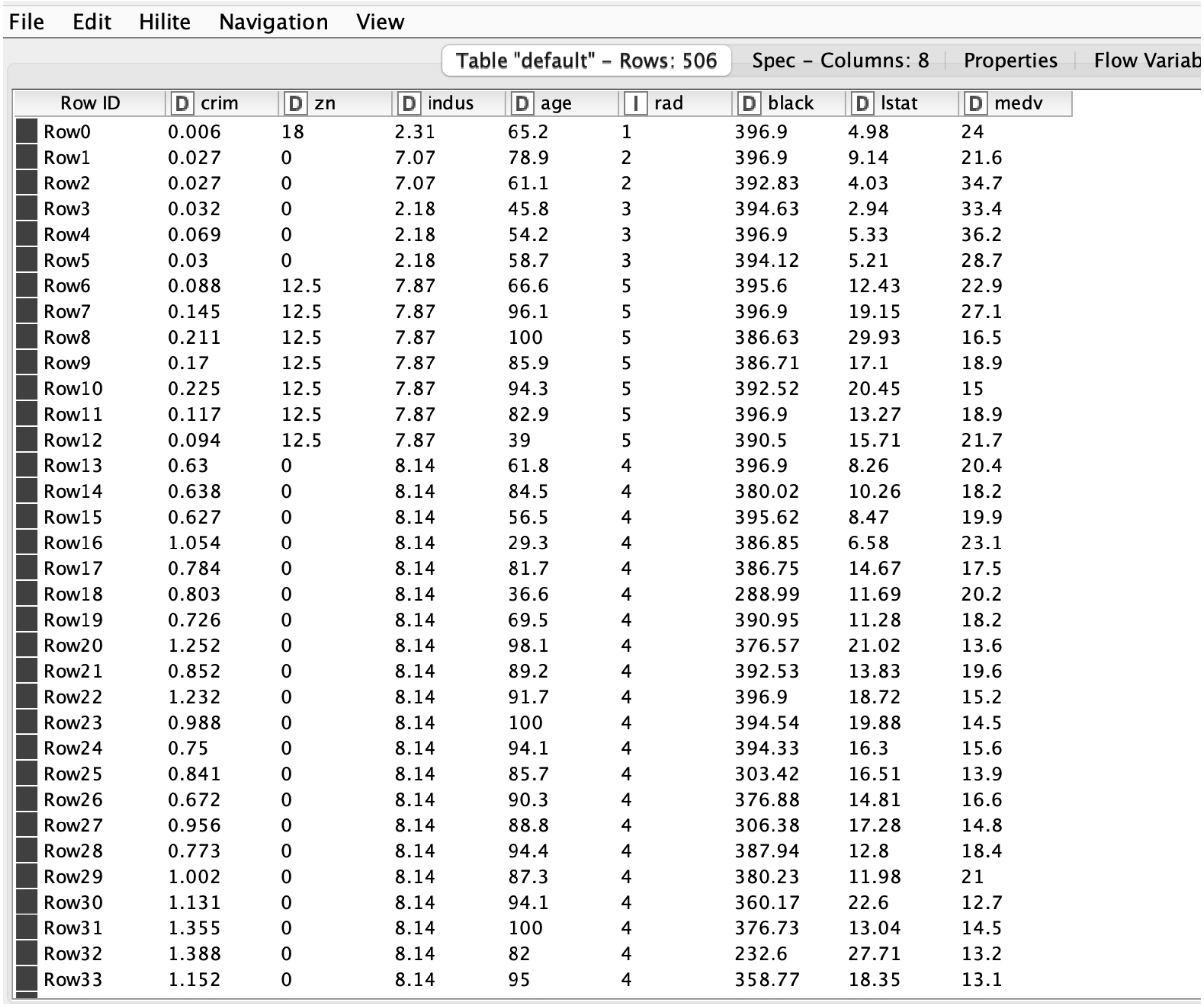
Their no missing value in the data. Therefore, no need for missing data filtering.

Low variance

Since the column with variance = 0, contain no useful information. Therefore, the some of the attributes are removed from the dataset.

High correlation

If two columns have correlation, then will contain the same information. Therefore, the columns with correlation more than 0.8 will contain the same information so one of them are excluded. The following feature are result of the process.



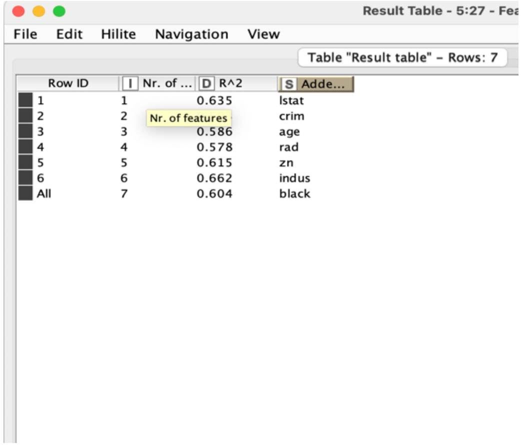
Row ID	D crim	D zn	D indus	D age	I rad	D black	D lstat	D medv
Row0	0.006	18	2.31	65.2	1	396.9	4.98	24
Row1	0.027	0	7.07	78.9	2	396.9	9.14	21.6
Row2	0.027	0	7.07	61.1	2	392.83	4.03	34.7
Row3	0.032	0	2.18	45.8	3	394.63	2.94	33.4
Row4	0.069	0	2.18	54.2	3	396.9	5.33	36.2
Row5	0.03	0	2.18	58.7	3	394.12	5.21	28.7
Row6	0.088	12.5	7.87	66.6	5	395.6	12.43	22.9
Row7	0.145	12.5	7.87	96.1	5	396.9	19.15	27.1
Row8	0.211	12.5	7.87	100	5	386.63	29.93	16.5
Row9	0.17	12.5	7.87	85.9	5	386.71	17.1	18.9
Row10	0.225	12.5	7.87	94.3	5	392.52	20.45	15
Row11	0.117	12.5	7.87	82.9	5	396.9	13.27	18.9
Row12	0.094	12.5	7.87	39	5	390.5	15.71	21.7
Row13	0.63	0	8.14	61.8	4	396.9	8.26	20.4
Row14	0.638	0	8.14	84.5	4	380.02	10.26	18.2
Row15	0.627	0	8.14	56.5	4	395.62	8.47	19.9
Row16	1.054	0	8.14	29.3	4	386.85	6.58	23.1
Row17	0.784	0	8.14	81.7	4	386.75	14.67	17.5
Row18	0.803	0	8.14	36.6	4	288.99	11.69	20.2
Row19	0.726	0	8.14	69.5	4	390.95	11.28	18.2
Row20	1.252	0	8.14	98.1	4	376.57	21.02	13.6
Row21	0.852	0	8.14	89.2	4	392.53	13.83	19.6
Row22	1.232	0	8.14	91.7	4	396.9	18.72	15.2
Row23	0.988	0	8.14	100	4	394.54	19.88	14.5
Row24	0.75	0	8.14	94.1	4	394.33	16.3	15.6
Row25	0.841	0	8.14	85.7	4	303.42	16.51	13.9
Row26	0.672	0	8.14	90.3	4	376.88	14.81	16.6
Row27	0.956	0	8.14	88.8	4	306.38	17.28	14.8
Row28	0.773	0	8.14	94.4	4	387.94	12.8	18.4
Row29	1.002	0	8.14	87.3	4	380.23	11.98	21
Row30	1.131	0	8.14	94.1	4	360.17	22.6	12.7
Row31	1.355	0	8.14	100	4	376.73	13.04	14.5
Row32	1.388	0	8.14	82	4	232.6	27.71	13.2
Row33	1.152	0	8.14	95	4	358.77	18.35	13.1

In the data analysis process a linear regression, Decision Tree regression, and polynomial regression Learner use for comparing two different criteria's (R-square and MSE).

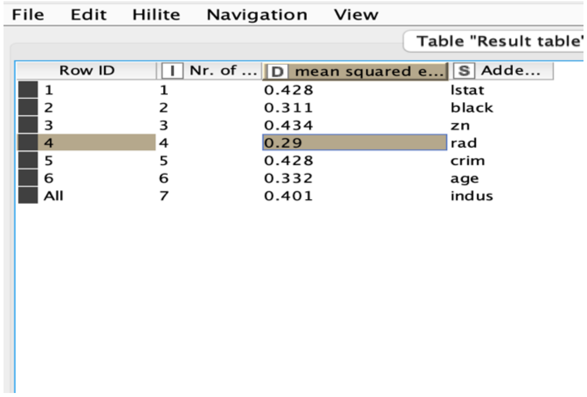
R-squared (R2) is a statistical measure that shows the amount of variance for the dependent variable explained by the independent variable or a regression model. It is a goodness-of-fit measure for regression models. When a regression model accounts for more of the variance, the data points are closer to the regression line.

While Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases.

The linear regression model has the following measure on the criteria's,



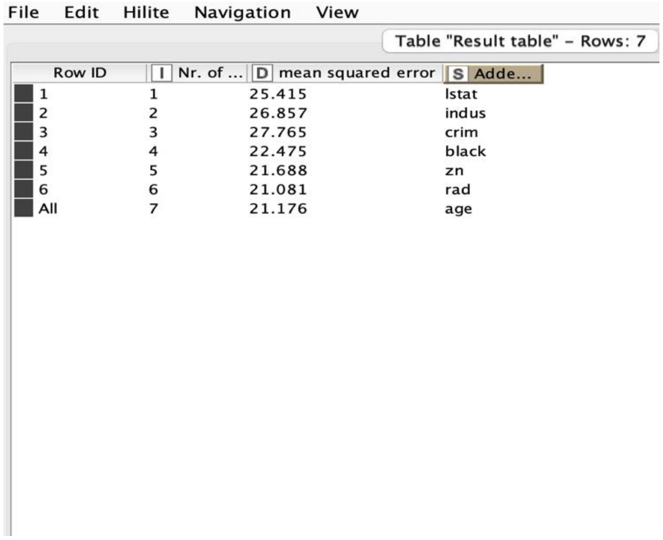
Row ID	Nr. of ...	R^2	Adde...
1	1	0.635	lstat
2	2		crim
3	3	0.586	age
4	4	0.578	rad
5	5	0.615	zn
6	6	0.662	indus
All	7	0.604	black



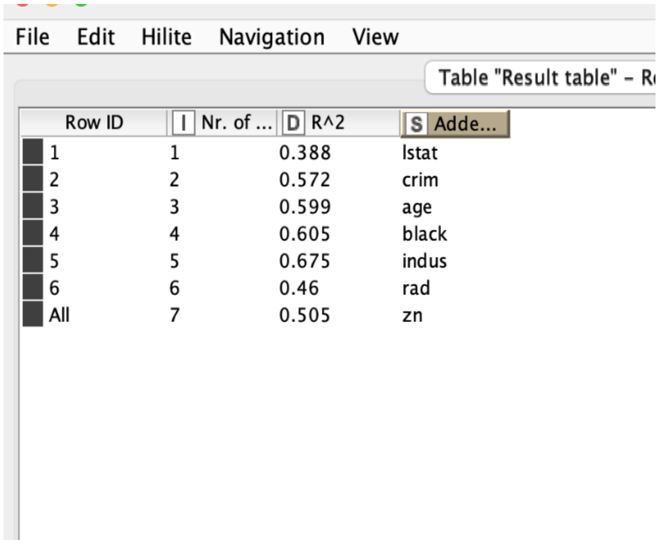
Row ID	Nr. of ...	mean squared e...	Adde...
1	1	0.428	lstat
2	2	0.311	black
3	3	0.434	zn
4	4	0.29	rad
5	5	0.428	crim
6	6	0.332	age
All	7	0.401	indus

As the figure depict the linear regression has 0.29 MSE and 0.662 R-square.

The Decision Tree regression model has the following measurement on the criteria's



Row ID	Nr. of ...	mean squared error	Adde...
1	1	25.415	lstat
2	2	26.857	indus
3	3	27.765	crim
4	4	22.475	black
5	5	21.688	zn
6	6	21.081	rad
All	7	21.176	age



Row ID	Nr. of ...	R^2	Adde...
1	1	0.388	lstat
2	2	0.572	crim
3	3	0.599	age
4	4	0.605	black
5	5	0.675	indus
6	6	0.46	rad
All	7	0.505	zn

As the figure shows the Decision Tree regression has 21.081 MSE and 0.675 R-square.

The polynomial regression model has the following measurement

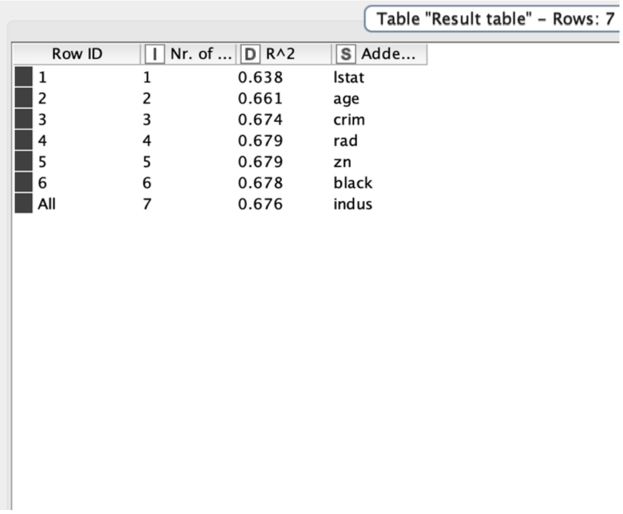


Table "Result table" – Rows: 7

Row ID	I Nr. of ...	D R^2	S Adde...
1	1	0.638	lstat
2	2	0.661	age
3	3	0.674	crim
4	4	0.679	rad
5	5	0.679	zn
6	6	0.678	black
All	7	0.676	indus

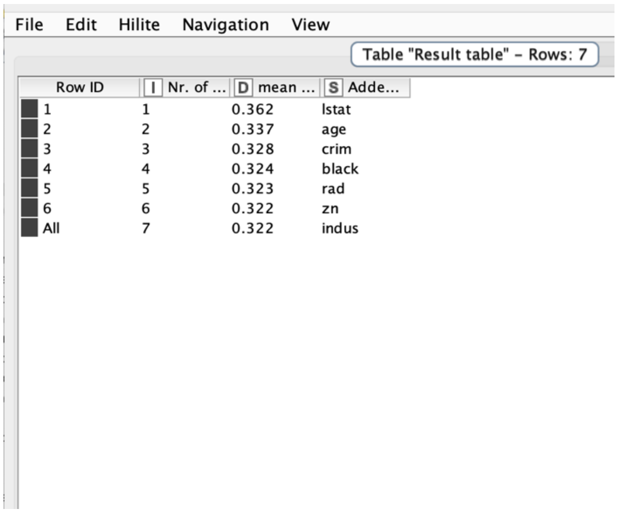


Table "Result table" – Rows: 7

Row ID	I Nr. of ...	D mean ...	S Adde...
1	1	0.362	lstat
2	2	0.337	age
3	3	0.328	crim
4	4	0.324	black
5	5	0.323	rad
6	6	0.322	zn
All	7	0.322	indus

The pictures show that the polynomial regression model has 0.322 MSE and 0.679 R-square and this value indicates that polynomial regression model is good compared to the other two models.