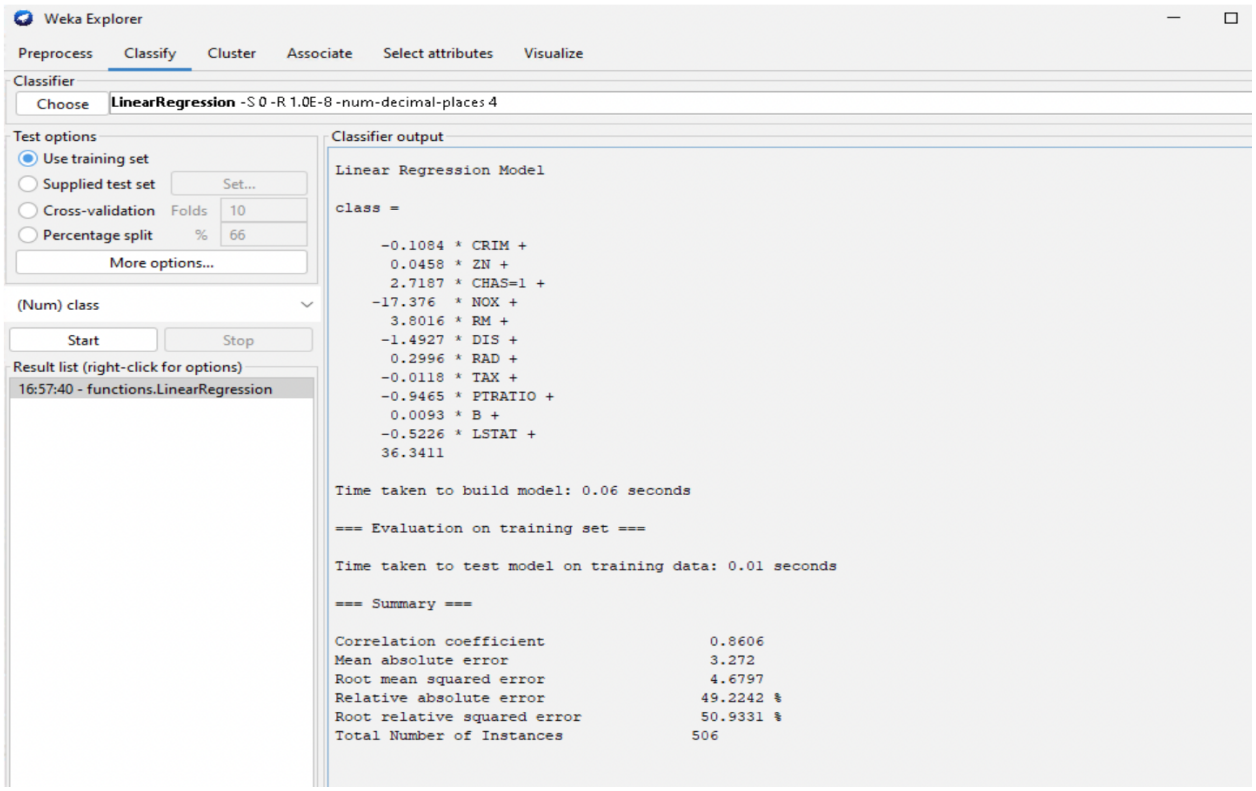
HW4: Data Mining - tool-based

Total points: 6

Question 1: WEKA

[1 point]
Screenshot with the linear regression model's equation.



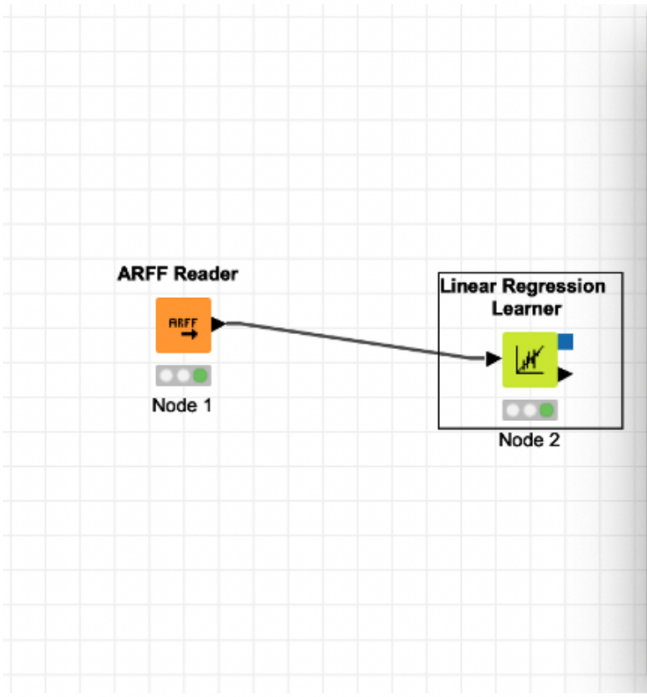
[1 point]
There are 12 terms in the equation (including the constant). It is acceptable if the constant is not included as a term and is mentioned separately. Each term affects the final MEDV value. A negative value would lower the median home price whereas a positive value raises it. (Also, explaining the M5 algorithm is fine - includes answers about Age/Indus)

Explanations that mean the same thing are accepted (e.g., terms like proportional and inversely proportional, explanation with examples, etc.) **For partially correct explanations, deduct 0.5**

Note: If a student got a different screenshot and number of variables, and they have mentioned the reasons for why they got those numbers, it is acceptable. Otherwise, If by any chance, the screenshots do not match (number of terms other than 12, different coefficients, different summary) that means the student has done it incorrectly; we can deduct 2 points.

Question 2: KNIME

[1 point]
Screenshot of the two nodes AARF Reader and Linear Regression Learner along with the model results/statistics. If either is missing/incorrect, deduct 0.5 points.



Linear Regression Result View - 3:2 - Line...

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
sex=I	-0.8249	0.1024	-8.0558	1.11E-15
sex=M	0.0577	0.0833	0.6925	0.4887
length	-0.4583	1.8091	-0.2533	0.8
diameter	11.0751	2.2273	4.9725	6.88E-7
height	10.7615	1.5362	7.0053	2.86E-12
whole_weight	8.9754	0.7254	12.373	0.0
shucked_weight	-19.7869	0.8174	-24.2086	0.0
viscera_weight	-10.5818	1.2937	-8.1792	4.44E-16
shell_weight	8.7418	1.1247	7.7723	9.55E-15
Intercept	3.8946	0.2916	13.3576	0.0

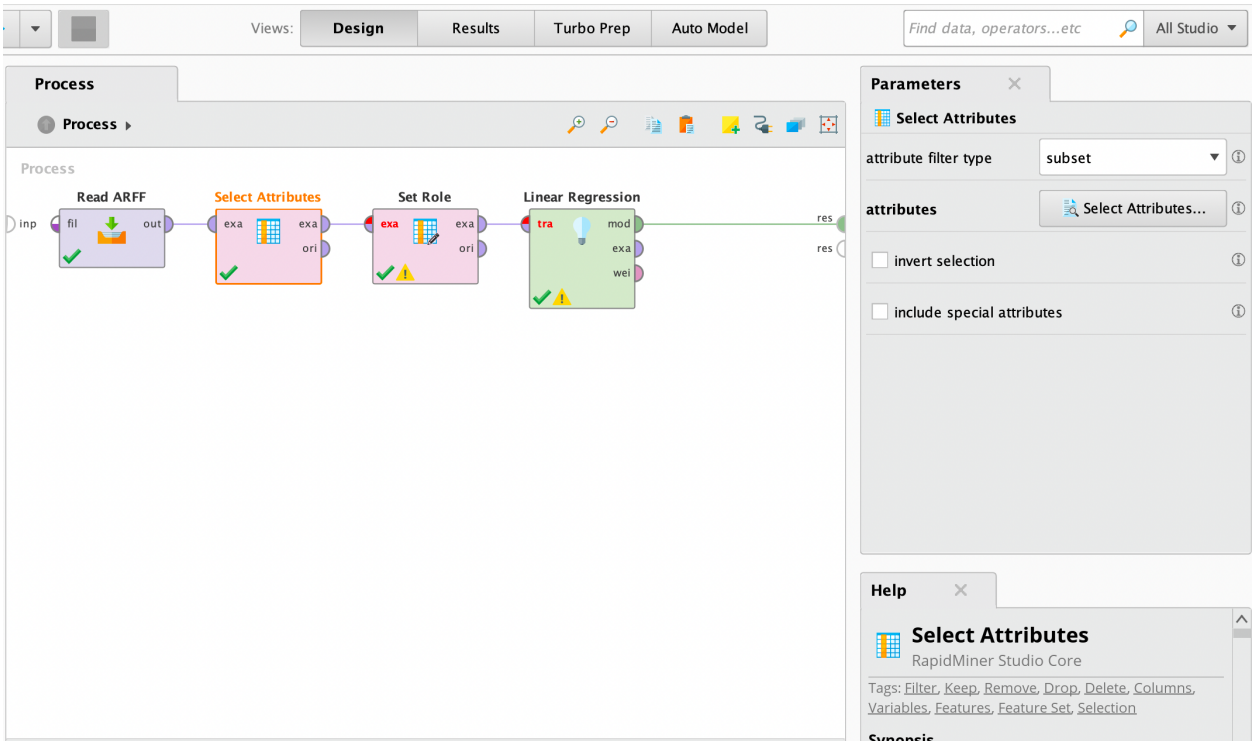
Multiple R-Squared: 0.5379
Adjusted R-Squared: 0.5369

[1 point]
README.txt file should contain the correct linear regression equation mentioned below.

```
num_rings = -0.8249 * sex = I +
            0.0577 * sex = M +
            -0.4583 * length +
            11.0751 * diameter +
            10.7615 * height +
            8.9754 * whole_weight +
            -19.7869 * shucked_weight +
            -10.5818 * viscera_weight +
            8.7418 * shell_weight +
            3.8946
```

Question 3: RapidMiner

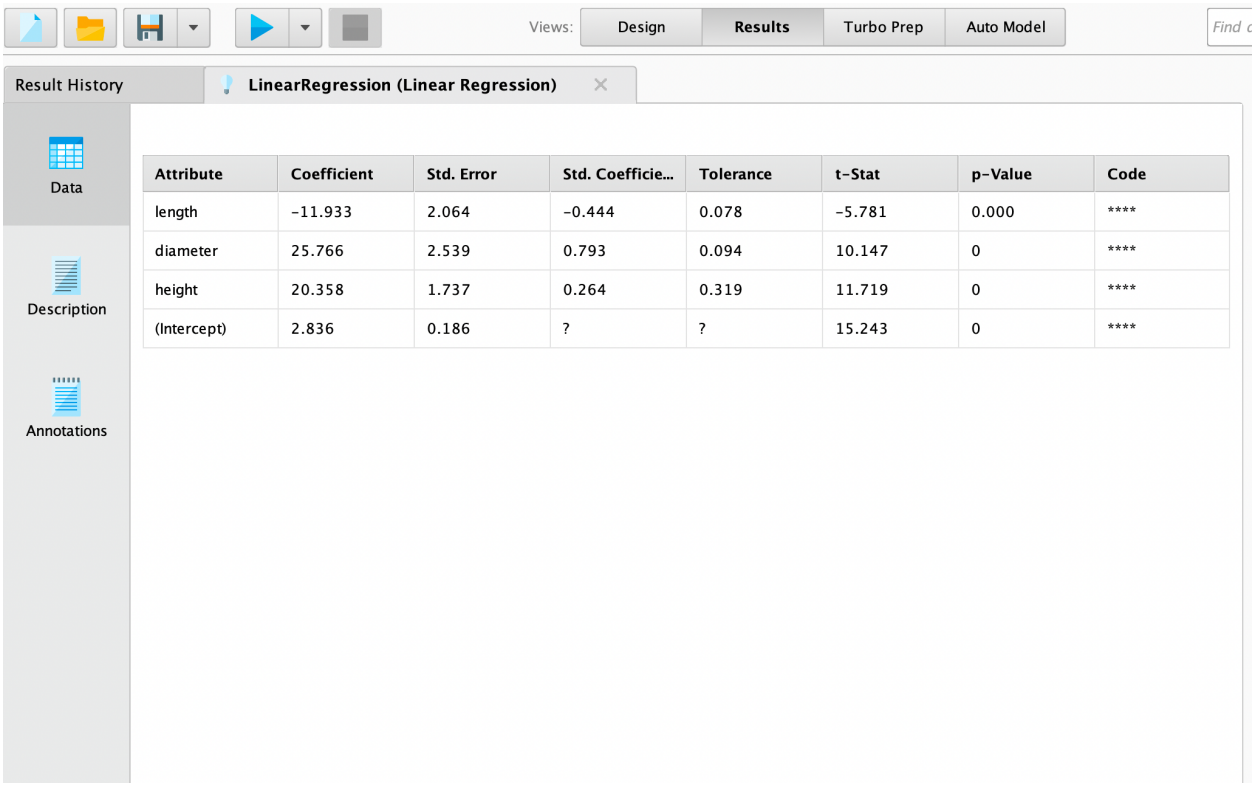
[1 point]
The screenshot of the expected design, as shown in the screenshot below. Deduct 0.25 points for each incorrect node.



[1 point]
README.txt file should contain the linear regression equation mentioned below.

```
- 11.933 * length
+ 25.766 * diameter
+ 20.358 * height
+ 2.836
```

For reference:



Attribute	Coefficient	Std. Error	Std. Coefficie...	Tolerance	t-Stat	p-Value	Code
length	-11.933	2.064	-0.444	0.078	-5.781	0.000	****
diameter	25.766	2.539	0.793	0.094	10.147	0	****
height	20.358	1.737	0.264	0.319	11.719	0	****
(Intercept)	2.836	0.186	?	?	15.243	0	****