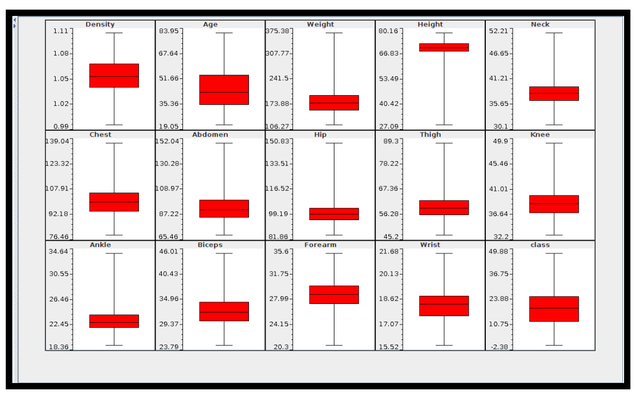
ADAMS: Visualization plug-ins

Michael Fowke: msf8@waikato.ac.nz
Peter Reutemann: fracpete@waikato.ac.nz
Geoff Holmes: geoff@scms.waikato.ac.nz

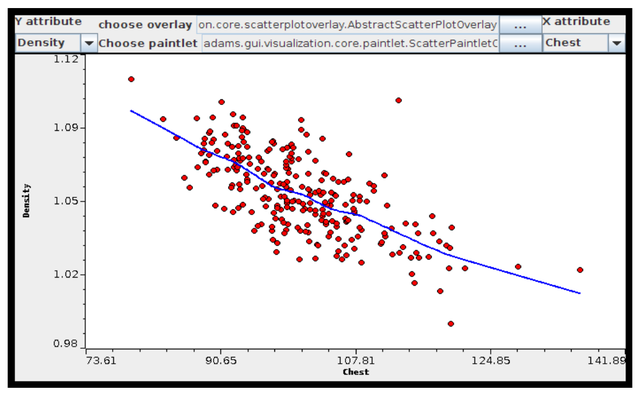
University of Waikato Summer Research Programme

Computer Science Department, University of Waikato, Private bag 3105, Hamilton, New Zealand

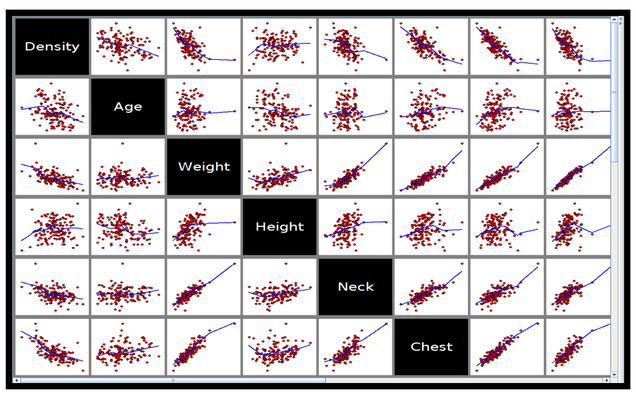
ADAMS, the Advanced Data mining and Machine learning System, provides a workflow-centric environment that allows researchers to set up and perform experiments. Popular machine learning packages, like WEKA [1] and MOA [2] are available to the user. By using the workflow approach, an experiment is broken down in (potentially many) individual data processing and evaluation steps, documenting it implicitly.



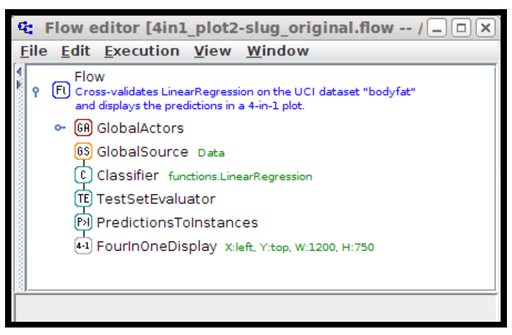
Box Plot visualization: Displays statistics for each attribute graphically.



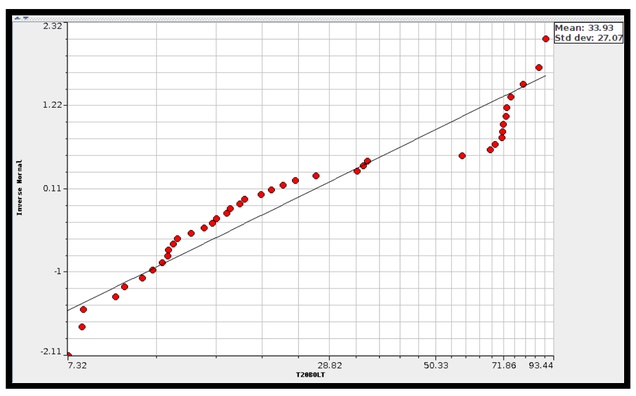
Scatter Plot visualization: Displays an attribute plotted against any other attribute with overlay and paintlet options available.



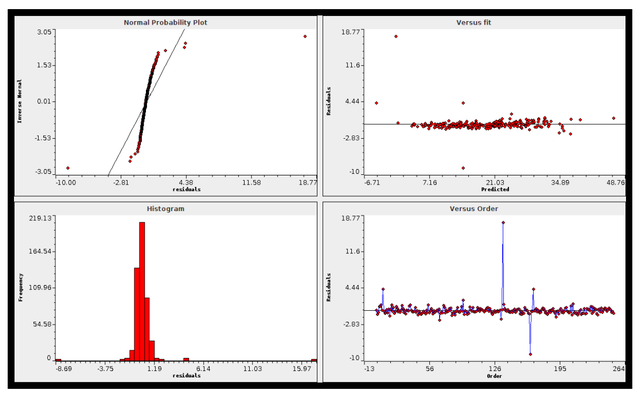
Matrix Plot visualization: Displays each attribute plotted against all other attributes.



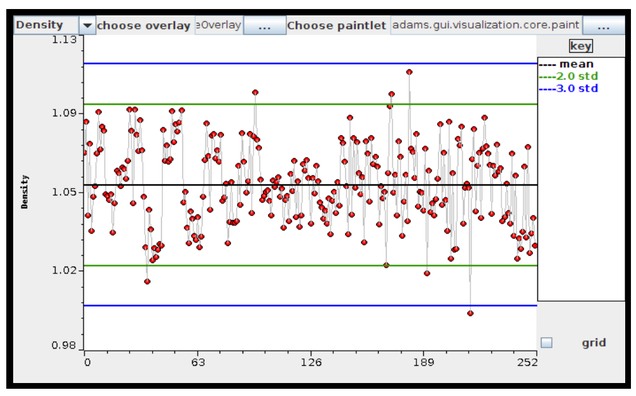
Flow editor: Used to define processing steps and set up visualizations.



Probability Plot visualization: Applies a regression to the dataset and transforms the data in an attempt to create a straight line. This illustrates a good regression fit.



Four-in-one visualization: Displays a normal probability plot, histogram, vs fit scatter plot and vs order scatter plot. Plotting the residuals from the classifier.



Z-Score visualization: Displays all instances for an attribute as well as standard deviation and mean overlays.

Why do we need visualizations?

Though researchers rely very much on summary statistics obtained from experiments to judge the validity and performance of built models, visualizing the results provides in many cases a more accurate picture of the performance. The new visualization plug-ins offer functionality commonly found in commercial statistical packages like Minitab [3].

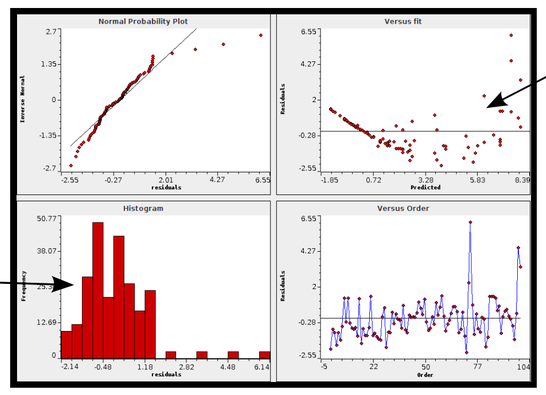
Example visualizations using slug dataset [4]

Correlation coefficient	0.9116
Mean absolute error	0.8864
Root mean squared error	1.2337
Relative absolute error	38.3116 %
Root relative squared error	41.1182 %
Total Number of Instances	100

Correlation coefficient appears alright

Errors are a bit high

Text output: Difficult to see that the model is not a good fit.



Non-random pattern, indicates a bad fit

Histogram not bell shaped

4-in-1 plot: More obvious that the model is not a good fit.

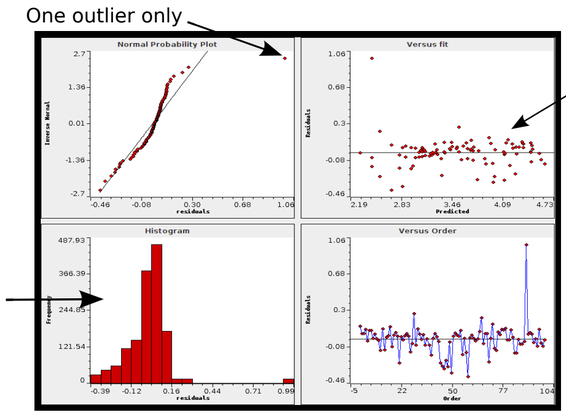
Take $\log_e$ of all the datapoints

Correlation coefficient	0.9694
Mean absolute error	0.093
Root mean squared error	0.1525
Relative absolute error	17.9266 %
Root relative squared error	24.5506 %
Total Number of Instances	100

Higher correlation coefficient

Errors are lower

Text output: Likely to be a better model.



One outlier only

Random pattern, indicates a good fit

Bell shaped histogram

4-in-1 plot: Easy to see this is a better fit.

References

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
[2] Albert Bifet, Geoff Holmes, Richard Kirkby, Bernhard Pfahringer (2010); MOA: Massive Online Analysis, Journal of Machine Learning Research (JMLR), Volume 11, 1601-1604
[3] Minitab 15 Statistical Software (2007). [Computer software]. State.College, PA: Minitab, Inc. (www.minitab.com)
[4] Barker, G, and McGhie, R (1984) The Biology of Introduced Slugs (Pulmonata) in New Zealand: Introduction and Notes on Limax Maximus, NZ Entomologist 8, pp 106-111