

1.1.1 Software Interface Review

Table 1 provides a breakdown of some of the key elements reviewed for each package in this study.

Table 1

Review Areas	Software				
	<i>Ctree</i>	<i>PolyAnalyst</i>	<i>See5</i>	<i>Cubist</i>	<i>Neuralyst</i>
Easy of Navigation	Very simple design, graphics and text based	The features were very complex that it was hard to understand	Simple interface most work done through text files though	Simple interface most work done through text files though	Features not explained in software, requires understanding NN terms and/or the manual
Graphical User Interface	Yes	Yes	Limited, to only model building	Limited, to only model building	Limited, through menus
Data Sources	Excel	Excel directly and various file formats	Text files formatted with another text file description	Text files formatted with another text file description	Excel
Help	Some built in help files	Extensive documentation and help tips	Some documentation	Some documentation	Extensive External documentation
Number of Methods for Analysis Available	1 almost 2, rules are generated but not able to implement.	11	2	1	1
“Black-Box”	Yes, however somewhat explained in documentation	Some, depends on the method used.	The rules are clear to understand, but how they are created is not clear with this version. Does allow extensive tweaking.	The rules are clear to understand, but how they are created is not clear. Does allow extensive tweaking.	Very much so, however allows some tweaking.
Model Parameter Options	Good, not too advanced (like See5) but straight forward	Good automatic model generation but does not for many advanced options	Many advanced options and some automatic options	Many advanced options and some automatic options	Extensive model control, however requires some advanced knowledge

1.1.2 Advantages and Disadvantages for Analysis Methods

Table 2 provides some of the key advantages and disadvantages for the methods previously used.

Table 2: Some Advantages and Disadvantages for Analysis Methods

Model Type	Advantages	Disadvantages
MDA	<p>Precise and clear conclusions</p> <p>Uniform</p> <p>Reliability can be statistically evaluated</p> <p>Faster and less costly than some tools</p> <p>Can quickly weed out extremes, allowing work with gray areas using other models</p> <p>Allows time-series data</p> <p>Generally used as a benchmark for studies, easy comparisons to previous studies</p> <p>Values used not influenced by opinion, because quantitative</p>	<p>Unable to use qualitative values</p> <p>No missing values</p> <p>Hard to set-up correctly, especially for beginners</p> <p>Problems with multi-collinearity</p> <p>Outliers can create problems</p> <p>Grouping variable must have limited distinct categories and be coded as integers (not continuous)</p> <p>Independent variables that are nominal must be recoded to dummy or contrast variables</p> <p>Multivariate normal distribution</p> <p>Only provides ordinal ranking</p> <p>Difficulty in determining the significance level of variables in the overall score</p> <p>Cases be independent</p> <p>Variance-covariance matrices be equal across the groups</p> <p>Members of each group only belong to one group and that all cases are part of one of the groups</p> <p>Data must be standardized</p>
ANNs	<p>Allows missing values</p> <p>Easy to set-up</p> <p>Offers qualitative elements</p> <p>Can discover cause-and-effect relationships and subtle patterns</p> <p>Continues to learn with new data</p> <p>Can learn with new data</p> <p>Shown to be good at classification</p> <p>Good at allowing variance because it allows generalizations</p> <p>Some good successes in predicting short-term results</p> <p>Can deal better with missing data, outliers, and multi-collinearity than many statistical methods</p> <p>Require little or no prior knowledge of the problem</p> <p>Is a feasible approach when processes are not well understood and theoretical models would be too complex or time consuming to create</p> <p>High tolerance for noisy data</p> <p>Sometimes can classify patterns not trained with</p> <p>Rule extraction programs have been</p>	<p>Not easy to add time-series data</p> <p>Overfitting can easily occur</p> <p>Hard to extract rules and reasons for predictions (not provide a clean model to a problem)</p> <p>“Black box” - poor interpretability</p> <p>Bad at math or when generalizations is not acceptable</p> <p>Less success for long-term results, partially because of the difficulty in adding the time-element</p> <p>Complex models often require hundreds of parameters and significant amounts of data for proper training, which can be very processor and time consuming</p> <p>Long training times</p> <p>Inputs mathematically related to the target</p> <p>If a test case is similar to a training case they should have similar targets</p> <p>Must be a large enough training set to cover the problem sufficiently</p>

	developed with some success	
Univariate Model	<p>Very easy to understand and easy to create</p> <p>Easy to evaluate</p> <p>Very quick to apply</p> <p>Cost effective</p> <p>Good for identifying outliers in data and initial data review</p>	<p>Do not capture enough variables, over simplifies complex problems</p> <p>Very limited to specific types of problems</p> <p>To be predictive requires an environment which will not adapt to the variables, removing symptoms</p> <p>Deciding which variables to use requires extensive testing</p> <p>The majority of negatives for MDA is also applicable here</p>
Logistic regression	<p>The exponentiated logistic coefficients can be interpreted as odds ratios</p> <p>More programs available with logistic regression than Probit</p> <p>Presence of multicollinearity will not lead to biased coefficients</p> <p>Can also be applied to ordered categories, however difficult to set-up</p> <p>Does not require normal distribution of variables</p> <p>Designed to find probability of inclusion in a group (known binomial distribution)</p> <p>Allows additivity</p> <p>If the categorical dependent reflects an underlying qualitative variable</p>	<p>Requires binary data in order to model probabilities of specified outcomes.</p> <p>Omitted variables can result in bias</p> <p>Inclusion of irrelevant variables can result in a poor model</p> <p>Errors in functional form can result in biased coefficient</p> <p>With multicollinearity the standard errors of the coefficients will be inflated</p> <p>Structural breaks in data</p> <p>Requires knowledge for how to set up</p> <p>Can become very complex</p> <p>Systematic values, not random variation</p> <p>No large outliers</p> <p>If continuous data are categorized information is lost and results may be misleading</p>
Probit	<p>Presence of multicollinearity will not lead to biased coefficients</p> <p>Can also be applied to ordered categories, however difficult to set-up</p> <p>Does not require normal distribution of variables</p> <p>Designed to find probability of inclusion in a group (known binomial distribution)</p> <p>Allows additivity</p> <p>If the categorical dependent reflects an underlying quantitative variable</p>	<p>The exponentiated logistic coefficients can not be interpreted as odds ratios</p> <p>Fewer programs available with Probit than Logit</p> <p>Must be a relation between the model parameters and the probit</p> <p>Assumes the categorical dependent reflects an quantitative variable and it uses a cumulative normal distribution</p> <p>Not recommended when there are many cases in either tail of a distribution</p> <p>Omitted variables can result in bias</p> <p>Inclusion of irrelevant variables can result in a poor model</p> <p>Errors in functional form can result in biased coefficient</p> <p>With multicollinearity the standard errors of the coefficients will be inflated</p> <p>Structural breaks in data</p> <p>Requires knowledge for how to set up</p> <p>Can become very complex</p> <p>Systematic values, not random variation</p>

		<p>No large outliers</p> <p>If continuous data are categorized information is lost and results may be misleading</p>
Recursive partitioning	<p>Software is easy to use</p> <p>Outcomes are easy to implement in a variety of systems</p> <p>Good for visual representation of the model</p> <p>Not sensitive to the data distribution</p> <p>Allows nonparametric data</p> <p>Can use categorical data and continuous data</p>	<p>The various algorithms (or their improvements) are often proprietary (which does not allow for testing and improvements from outside)</p> <p>Hard to update with new data</p> <p>Overfitting can occur</p> <p>Fragmentation</p> <p>Repetition</p> <p>Replication</p> <p>Scalability issues for large datasets due to memory requirements and processing time, however new methods show promise</p> <p>Many splits to deal with outliers</p>
Expert Systems	<p>Offers qualitative methods</p> <p>Easy to modify</p> <p>Good when human expertise is not always available when needed</p> <p>Provides consistency with problem handling</p> <p>Allows tracing steps</p>	<p>Requires an expert</p> <p>Experts have problems explaining their decision process (gut feelings, etc)</p> <p>Uses often imperfect knowledge to reach solution (heuristic)</p> <p>Very specific</p> <p>Can not learn</p> <p>Time-consuming and expensive to create</p> <p>Domain must be well structured</p> <p><i>NOT pattern or model discovery system (unlike previous methods) MUST be programmed with knowledge from an expert!</i></p>