**INTRODUCTION TO DECISION TREES**

by Simon Moss

|  |
| --- |
| **Introduction** |

**Machine learning**

Researchers often utilize an approach called machine learning to predict some outcome—such as whether someone is likely to complete their PhD—from a range of measures called inputs. These inputs might include grade point average, number of research papers, and proficiency in English, as measured by the test called IELTS.

Machine learning comprises a range of techniques, such as decision trees, random forests, neural networks, and support vector machines. This document is confined to decision trees. Decision trees actually include many variants, such as CHAID, CART, and QUEST.

**Regression analysis**

As discussed in a separate document, called Introduction to Neural Networks, statisticians often utilize another approach, rather than machine learning, to predict some outcome from a series of inputs. Specifically, they might utilize regression analysis—a technique that generates an equation that researchers can use to predict an outcome. The following box presents an example of an equation that regression analysis might generate

|  |
| --- |
| Likelihood of completion = .03 x GPA + .17 x no. of research paper + .03 x IELTS score |

For example, if someone arrives with a 4.5 GPA, 2 research papers, and a 6.5 IELTS score—a measure of proficiency in English

* Likelihood of completion = .03 x 4.5 + .17 x 2 + .03 x 6.4 = .409
* So, the likelihood this person will complete is about .41 or 41%

**Problems with regression**

Unfortunately, in many circumstances, these simple equations are not sufficient to explain the association between the inputs and outcome. For example

* perhaps GPA is strongly related to completion, but only when number of papers is below 2 and IELTS is above 6.5 but below 7.5
* perhaps number of papers is strongly related to completion, but only when GPA is between 4 and 5
* hence, the association between the predictors and outcomes might depend on the precise value of other predictors

Even sophisticated variants of regression analysis—including non-linear regression, linear mixed models, multinomial logistic regression, or even generalized estimating equations—disregard these complications. The inclusion of interaction terms, sometimes called moderated regression, is helpful but not usually sufficient to accommodate all the ways in which the values on one predictor or input may affect the impact of other predictors or inputs.

**Role of machine learning**

Instead, another branch of analysis, called machine learning, is preferable in this instance. This document, however, will be confined to a subset of machine learning called decision trees. Decision trees are also called hierarchical splitting, segmentation, and other names.

|  |
| --- |
| **Decision trees: General principles** |

**Simple illustration**

To illustrate the concept of decision trees, suppose researchers want to predict which candidates are likely to complete their PhD or Masters by Research. Initially, the researchers will collect data about previous applicants. An extract of these data appear in the following table.

**Overall**: 60% completions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Person | Did this person complete | Highest degree | Published papers | IELTS score |
| John | Yes | Research Masters | Yes | 6 |
| Karen | Yes | Research  Masters | No | 6 |
| Len | No | Honours | No | 5 |
| Marsha | No | Honours | Yes | 6 |
| Neil | Yes | Bachelor | No | 8 |
| … | … | … | … | … |
| Olivia | No | Coursework Masters | No | 6 |

In essence, the rationale is that some algorithm splits each input—such as highest degree, published papers, or IELTS score—into subsets. For example

* the algorithm might divide highest degree into two subsets: Research Masters versus other degrees
* the algorithm then determines whether each subset is pure—that is, whether everyone in this subset generated the same outcome
* in this instance, everyone who has attained a Research Masters —the darker shading—has completed their PhD. So, this subset is pure
* But, the subset of remaining individuals are not pure, as the right side of the following diagram shows. That is, of the people who have attained another degree, at least one person completed, and at least three people did not complete, their thesis.



* If a subset is not pure, the algorithm then splits these individuals on another input, such as IELTS above 6.5 versus IELTS below 6.5
* The following diagram illustrates this outcome



In this instance—for people who have not completed a Research Masters —everyone with an IELTs above 6.5 completed their thesis. Everyone with an IELTS below 6.5 did not complete their thesis. These subsets are pure. Consequently

* In the future, if an applicant has attained a Research Masters, we can be confident they will complete their thesis
* If an applicant has not attained a Research Masters, but scored an IELTS above 6.5, we can also be confident they will complete their thesis
* Otherwise, we can be confident the individuals will not complete their thesis

In practice, however, decision trees are not as straightforward. More branches might be needed to generate pure subsets.

**But how does the algorithm decide which input to split?**

In the previous example, the algorithm first splits people into individuals who had completed a Research Masters and individuals who had not completed their Research Masters. Then, for the latter category, the algorithm splits people into individuals whose IELTS exceeded 6.5 and individuals whose IELTS was lower than 6.5. The question, then, is how does the algorithm decide which inputs or characteristics to split first. In short

* Researchers have developed a range of criteria to decide the order in which the inputs are split
* In essence, the algorithm identifies inputs that are most related to the outcomes.
* For example, in the previous example, highest education was highly associated with completion—and thus split first
* Number of papers was not highly associated with completion—and, therefore, was not split first.

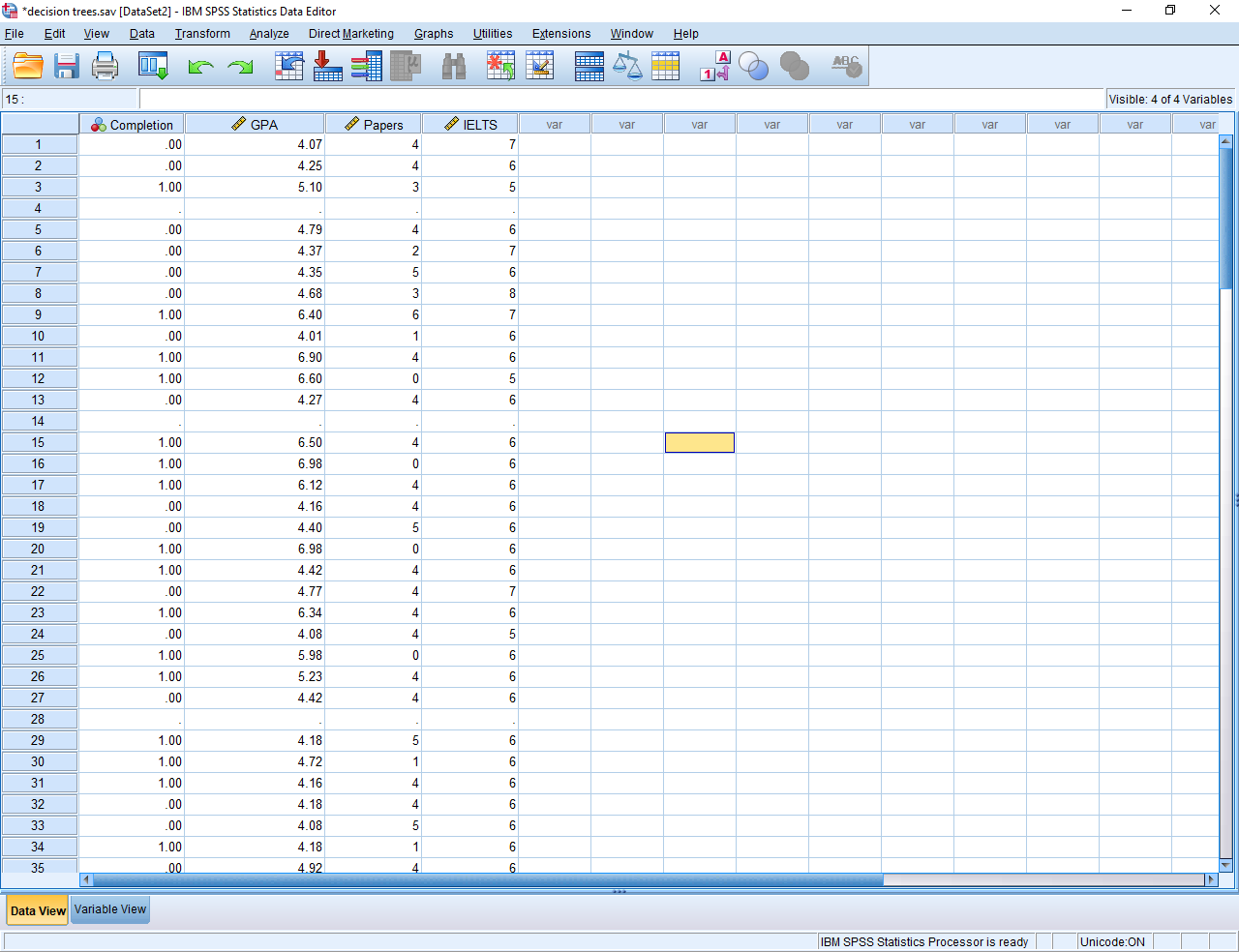
For example, one variant of this approach is called CHAID—or chi square automatic interaction detector. In essence, at each point in the tree,

* The algorithm calculates a chi-square test of independence for each input or attribute.
* This test determines the extent to which the outcome—completion—depends on the input or attribute
* A low p values indicates the outcome greatly depends on the input
* So, the algorithm choose the input that generates the lower p value
* This procedure is repeated at each point in the tree, sometimes called nodes

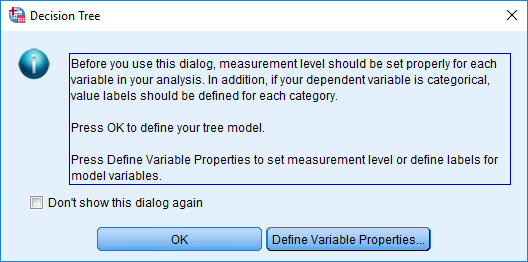
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outcome | Research Masters | Not Research Masters |  | Papers | No papers |
| Completed | 8 | 2 |  | 7 | 9 |
| Not completed | 2 | 6 |  | 4 | 5 |
|  | p = | .04 |  | p = | .24 |

|  |
| --- |
| **Examples with SPSS** |

To illustrate how researchers can utilize software to generate these decision trees, consider this example, with SPSS. Even if you do not use SPSS, this example could be useful. In particular, to conduct decision trees in SPSS, first enter the data. In this instance, each row corresponds to a separate person or unit. Each column corresponds to a separate characteristic, such as each input and the outcome.

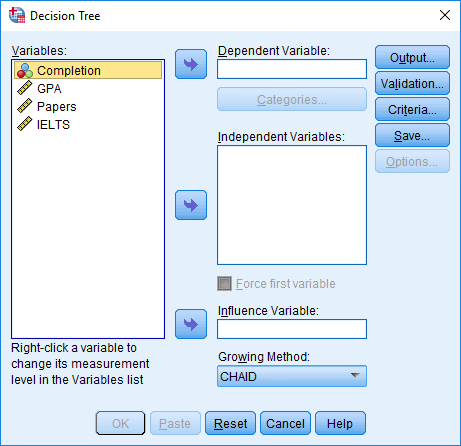


Then, in the *Analyze* menu, choose *Classify* and then *Tree*. These choices will sometimes generate the following screen.

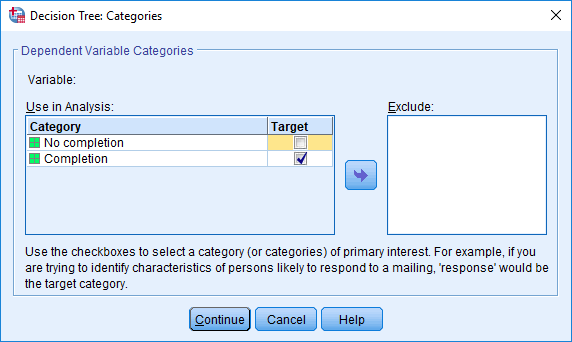


This screen is designed to remind you to define your variables properly in the *Variable* *View tab.* For example

* In the column called *Measures,* numerical variables should be designated as *Scale*. Categorical variables should be designated as *Nominal*.
* In the column called *Values*, assign a label to each category of the categorical variables
* Once you press OK, the following screen will appear



* In the box labelled *Dependent Variables*, choose the outcome you want to predict, such as *Completion*
* In the box labelled *Independent Variables*, choose the inputs*.* You can include many characteristics, because the algorithm will discard redundant inputs
* In the box labelled *Growing Method*, you can choose a range of options, such as *CHAID*. These alternatives will be discussed later.
* Press Categories to generate the next screen

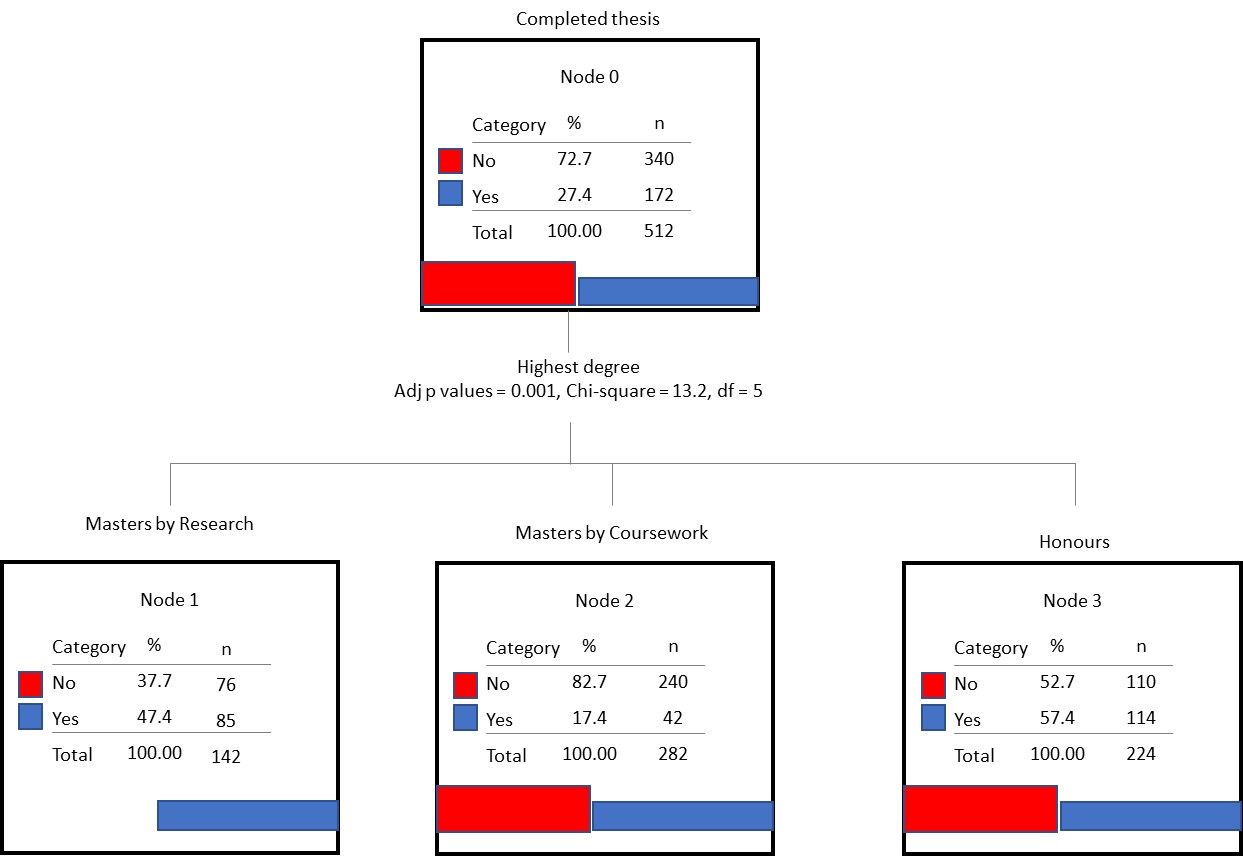


* Tick the category that represents your main interest or priority.
* For example, in this instance, perhaps the researcher is most interested in the candidates who have completed their research. This decision is quite arbitrary however.
* Then press *Continue* and *OK*.

**How to interpret the results**

After a while, SPSS will generate a decision tree, similar to the previous example, as well as some other output. Here is an extract of the tree. To illustrate

* The top box, called Node 1, specifies the number of individuals who have not completed and completed their thesis
* The chi-square test below this box indicates that completion is highly associated with highest education, because the p value is low
* The next set of boxes present the completion rate associated with each degree, and so forth
* Typically, these trees might comprise 3 to 5 levels



|  |
| --- |
| **Key decisions** |

To construct decision trees, researchers need to reach several decisions. This section offers some insights that could help you reach these decisions

**Which growing method should you utilize?**

You need to decide which growing method or algorithm to use to split the datasets, such as CHAID, exhaustive CHAID, CRT, and QUEST. Here are some insights that can help you distinguish these alternatives.

* CHAID often splits the data into more than two categories, such as Masters by Research, Masters by Coursework, and Honours. You may or may not prefer this option. QUEST and CRT, also called CART or Classification and Regression Trees, tends to split the data into two categories
* CHAID is useful when the outcome is categorical. CART is useful regardless of whether the outcome is categorical or numerical
* Because CHAID relies on chi-square test of independence, the sample size needs to be quite large—perhaps more than about 100 for each category

**Generalizability**

One concern with decision trees is the results are unstable. Even minor changes to your sample can generate entirely different trees. Some more advanced techniques, such as random forests, bagging, or boosting, are designed to overcome this problem to some extent.

To assess the stability of your results, you could construct the decision tree with a subset of data—and then test the decision tree with the rest of your data. For example

* you might construct the decision tree with 70% of the data, called the training data
* you could then assess this decision tree with the remaining 30% of your data, called the hold-out sample
* to achieve this goal, before pressing *OK,* selecy *Validation* and choose *Split-sample validation*
* you can then specify the percentage of data you would like to use to construct the decision tree—a figure that often ranges from 50% to 75%
* this option generates two decision trees—from the testing sample and hold-out sample respectively
* the classification table specifies the accuracy of predictions for the testing sample and hold-out sample respectively

|  |
| --- |
| **Decision trees versus neural networks** |

Whenever you want to use machine learning, you can choose a variety of alternatives. Two of the most common techniques, especially for researchers who are not experts in this field, are decision trees and neural networks. This section helps you decide which of these approaches to apply in various circumstances. Nevertheless, some researchers utilize both approaches and then choose the more effective outcome.

|  |  |
| --- | --- |
| Benefits of decision trees | Benefits of neural networks |
| You can **interpret** decision trees more readily than neural networks. That is, after scanning the output, you might be able to conclude that “English is not as important if the individuals had excelled in all their degrees”. You cannot as readily derive conclusions from neural networks. | Unlike neural networks, decision trees may entail **assumptions**, such as axis-parallel splits of the data, that may not be accurate; neural networks are thus suitable in a broader range of circumstances, such as nonlinear interactions |
| Decision trees often, but not always, generate the algorithm more **rapidly** than neural networks; that is, they train faster | You can gradually improve the algorithm or network incrementally whenever you receive additional data; in contrast, if you construct decision trees, you need to use all the data during the training phase. |
| Once you have developed an algorithm, decision trees classify or predict outcomes more rapidly than neural networks because they exclude unnecessary inputs or predictors |  |