Decision tree analysis in SPSS

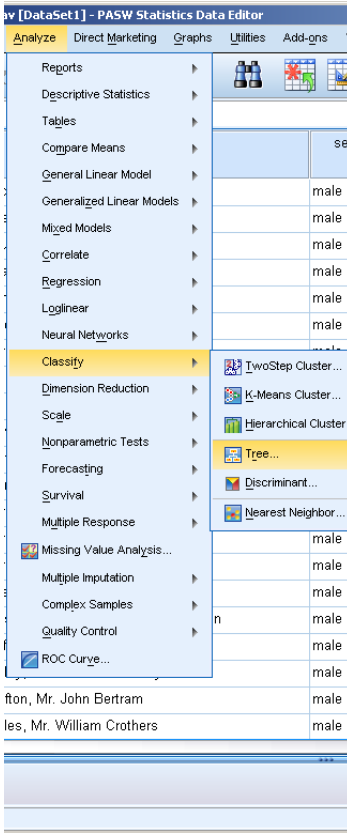
Maths and Statistics Help Centre

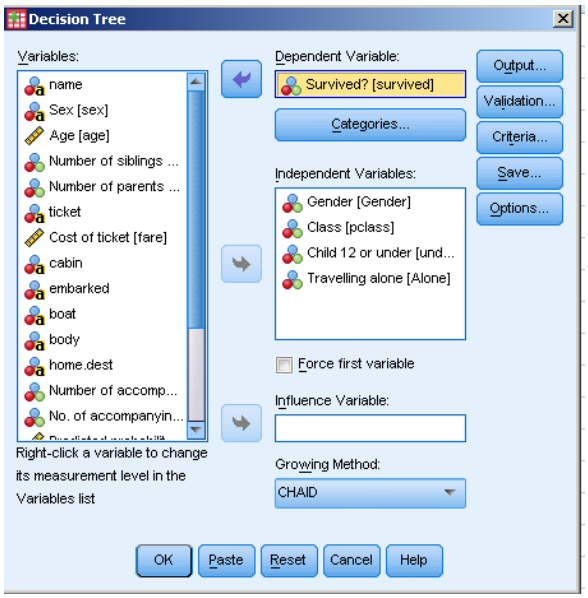
Introduction

Decision tree analysis helps identify characteristics of groups, looks at relationships between independent variables regarding the dependent variable and displays this information in a non-technical way. The process can also be used to identify classification rules for future events e.g. identifying people who are likely to belong to a particular group.

Basic model

The following example uses records from the Titanic on passengers. The tree will look at what factors affected chances of survival.





Dependent variable: Binary indicator of survival (1 = survived)

Independent variables:
Gender
Class (1^{st}, 2^{nd}, 3^{rd})
Child under 13 (Under 13, adult)
Travelling alone/ travelling with others.

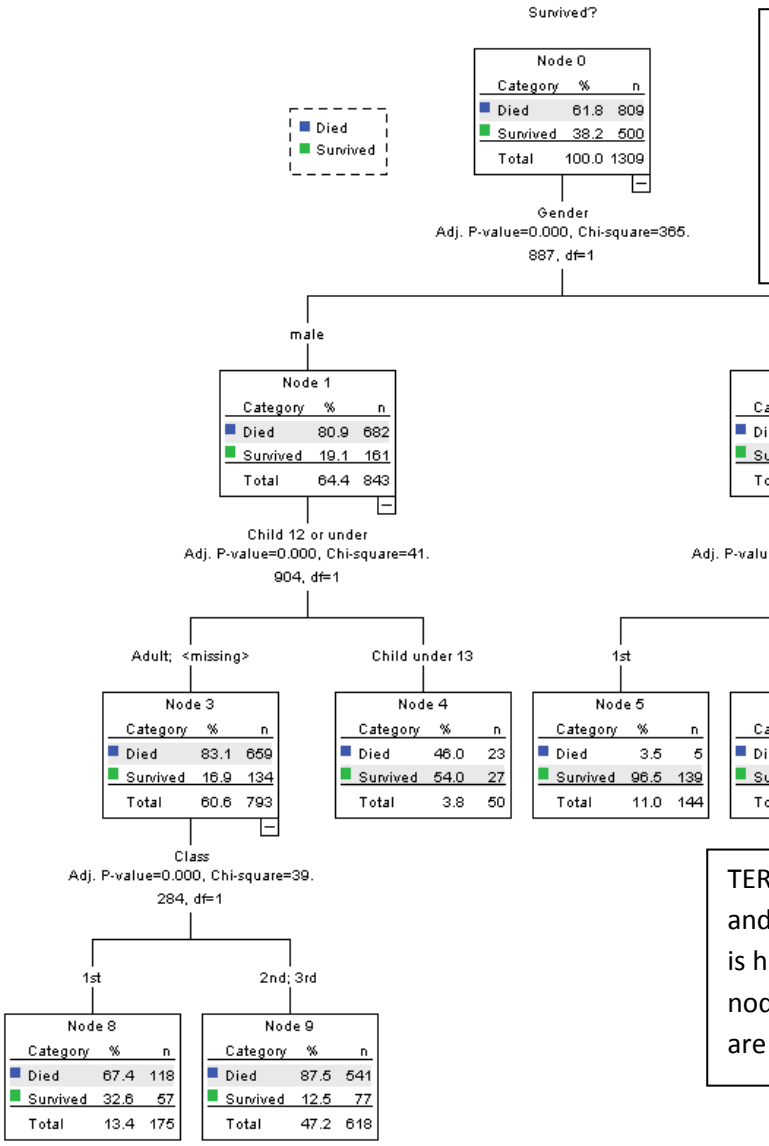
Growing method: The most commonly used growing methods are CHAID (Chi-squared automatic interaction detection) and CRT (Classification and regression).
Summary of differences:

- Treatment of missing values. CRT uses surrogates (classification via other independent variables with a high association with the independent variable with a missing value) whereas CHAID treats all missing values within an independent variable as one category.
- CHAID uses Pearson's Chi-squared to decide on variable splits and CRT uses Gini
- CRT only produces binary splits. If all independent variables are binary, the resulting tree from CRT and using the Pearson's Chi-squared option within CHAID will produce the same tree.
- CRT has a pruning ability so that extra nodes which do not increase the risk (wrong classification) by much can be automatically removed to leave a simpler tree.

Decision tree analysis in SPSS

Maths and Statistics Help Centre

Basic output using CHAID



For each node, the number of people and % who died/-survived is given. The splits occur in order of importance. Here, gender was the most significant factor regarding survival so the 'parent' node containing all 1309 passengers splits into two 'child' nodes, one containing males and the other females.

TERMINAL NODE: This is the end of a branch and a point of classification. The classification is highlighted in grey. For example, those in node 7 are classified as dying. These people are female and were in 3rd class.

Terminal node	Path	Classification	Number correct	Number wrong
4	Male →under 13	Survived	27	23
5	Female → 1st Class	Survived	139	5
6	Female → 2nd Class	Survived	94	12
7	Female → 3rd Class	Died	110	106
8	Male →Adult →1st Class	Died	118	57
9	Male →Adult →2nd or 3rd Class	Died	541	77

The risk represents the proportion of cases misclassified by the proposed classification. The classification table summarises the percentages classified correctly. The model classified 95.1% of those dying correctly, but only 52% of those who survived.

Decision tree analysis in SPSS

Maths and Statistics Help Centre

Risk

Estimate	Std. Error
.214	.011

Growing Method: CHAID
Dependent Variable: Survived?

Classification

Observed	Predicted		
	Died	Survived	Percent Correct
Died	769	40	95.1%
Survived	240	260	52.0%
Overall Percentage	77.1%	22.9%	78.6%

Growing Method: CHAID
Dependent Variable: Survived?

```
/* Node 1 */.
IF (((Gender = "male") OR (Gender != "female")  AND  (Number of accompanying siblings
or spouses != "1")))
THEN
Node = 1
Prediction = 0
Probability = 0.809015

/* Node 5 */.
IF (((Gender = "female") OR (Gender != "male")  AND  (Number of accompanying siblings
or spouses = "1")))  AND  (((Class = "1st" OR Class = "2nd") OR (Class != "3rd")  AND
((Age NOT MISSING  AND  (Age > 23.5)) OR Age IS MISSING  AND  (Number of accompanying
siblings or spouses != "3 or more"))))  AND  (((Class = "1st") OR (Class != "2nd")  AND
(Age IS MISSING  OR (Age > 34.5))))
THEN
Node = 5
Prediction = 1
Probability = 0.965278

/* Node 6 */.
IF (((Gender = "female") OR (Gender != "male")  AND  (Number of accompanying siblings
or spouses = "1")))  AND  (((Class = "1st" OR Class = "2nd") OR (Class != "3rd")  AND
((Age NOT MISSING  AND  (Age > 23.5)) OR Age IS MISSING  AND  (Number of accompanying
siblings or spouses != "3 or more"))))  AND  (((Class = "2nd") OR (Class != "1st")  AND
(Age NOT MISSING  AND  (Age <= 34.5))))
THEN
Node = 6
Prediction = 1
Probability = 0.886792

/* Node 4 */.
IF (((Gender = "female") OR (Gender != "male")  AND  (Number of accompanying siblings
or spouses = "1")))  AND  (((Class = "3rd") OR (Class != "1st"  AND  Class != "2nd")
AND  ((Age NOT MISSING  AND  (Age <= 23.5)) OR Age IS MISSING  AND  (Number of
accompanying siblings or spouses = "3 or more"))))
THEN
Node = 4
Prediction = 0
Probability = 0.509259
```

Decision tree analysis in SPSS

Maths and Statistics Help Centre