Machine Learning Applied to Fourteen Agricultural Datasets

Kirsten Thomson
Robert J. McQueen

Machine Learning Group
Department of Computer Science
University of Waikato

Research Report

September, 1996

Table of Contents

1. INTRODUCTION 1
1.1 Machine Learning 1
1.2 The WEKA machine learning workbench 2
1.3 Characteristics of the datasets 3
1.4 Classes and accuracy 3
1.5 WEKA Workbench functions and machine learning schemes used 4
1.6 The results of the tests 4
2. APPLE BRUISING 6
2.1 Original research 6
2.2 Machine learning 6
2.3 Discussion of results 12
2.4 Conclusion 12
3. BULLS 13
3.1 Original research 13
3.2 Machine learning 13
3.3 Discussion of results 15
3.4 Conclusion 15
4. EUCALYPTUS SOIL CONSERVATION 17
4.1 Original research 17
4.2 Machine learning 17
4.3 Discussion of results 25
4.4 Conclusion 25
5. GRASS GRUBS AND DAMAGE RANKING 30
5.1 Original research 30
5.2 Machine learning 30
5.3 Discussion of results 33
5.4 Conclusion 34
6. GRASS GRUBS AND RAINFALL 38
6.1 Original research 38
6.2 Machine learning 39
6.3 Discussion of results 41
6.4 Conclusion 42
7.GROWER QUESTIONNAIRE 46
7.1 Original research 46
7.2 Machine learning 46
7.3 Discussion of results 49
7.4 Conclusion 49
8. PASTURE PRODUCTION 51
8.1 Original research 51
8.2 Machine learning 51
8.3 Discussion of results 55
8.4 Conclusion 55
9. PEA SEED COLOUR 58
9.1 Original research 58
9.2 Machine learning 58
9.3 Conclusion 61

10. SLUGS 63
10.1 Original research 63
10.2 Machine learning 63
10.3 Discussion of results 68
10.4 Conclusion 68
11. SQUASH HARVEST 71
11.1 Original research 71
11.2 Machine learning 71
11.3 Conclusion 78
12. VALLEY CATEGORIES 83
12.1 Original research 83
12.2 Machine learning 83
12.3 Discussion of results 87
12.4 Conclusion 87
13. VENISON BRUISING 89
13.1 Original research 89
13.2 Machine learning 89
13.3 Discussion of results 95
13.4 Conclusion 95
14. WASP NESTS 97
14.1 Original research 97
14.2 Machine learning 97
14.3 Discussion of results 102
14.4 Conclusion 102
15. WHITE CLOVER PERSISTENCE TRIALS 103
15.1 Original research 103
15.2 Machine learning 103
15.3 Discussion of results 111
15.4 Conclusion 111
16. REFERENCES 114

1. INTRODUCTION

This document reports on an investigation conducted between November, 1995 and March, 1996 into the use of machine learning on 14 sets of data supplied by agricultural researchers in New Zealand. Our purpose here is to collect together short reports on trials with these datasets using the WEKA machine learning workbench, so that some understanding of the applicability and potential application of machine learning to similar datasets may result.

We gratefully acknowledge the support of the New Zealand agricultural researchers who provided their datasets to us for analysis so that we could better understand the nature and analysis requirements of the research they are undertaking, and whether machine learning techniques could contribute to other views of the phenomena they are studying. The contribution of Colleen Burrows, Stephen Garner, Kirsten Thomson, Stuart Yeates and James Littin and other members of the Machine Learning Group in performing the analyses was essential to the completion of this work.

This work is supported by a PGSF grant from the Government of New Zealand

1.1 Machine Learning

Machine learning is an emerging technology that can aid in the discovery of rules and patterns in sets of data. It has frequently been observed that the volume of recorded data is growing at an astonishing rate that far outstrips our ability to make sense of it, and the phrase “database mining” is now being used to describe efforts to analyze data sets automatically for significant structural regularities (Piatetsky–Shapiro and Frawley, 1991). Potential applications of these techniques in domains such as agriculture and horticulture are legion. There are many possible ways to capitalize on any patterns that are discovered. For example, their implicit predictive ability of decision and classification rules generated by machine learning techniques could be embedded in automatic processes such as expert systems, or they could be used directly for communication between human experts agricultural practitioners. The decision tree output of some machine learning schemes can aid in the discovery of the structure and hierarchy of observed data, assisting in the focusing of analysis efforts, and potential significant saving of time.

Machine learning has potential in assisting agricultural researchers to better understand the data that is the product of their field trials (McQueen et al, 1995). In the 14 cases discussed in this report, the data was collected by the original researchers anticipating the use of one or more statistical analysis techniques to understand its meaning. We anticipate a future time when machine learning might be one of the techniques used, alongside statistical analysis, to derive understanding from field data. If that is to be a common occurrence, then experiments and field data collection may need to anticipate the benefits of machine learning, and the requirements this may impose, or opportunities presented on how the study is designed and data captured.

1.2 The WEKA machine learning workbench

One of the practical problems in applying machine learning is that it is hard to acquire a variety of learning tools and experiment with them in a uniform way. The WEKA (Waikato Environment for Knowledge Analysis) workbench is a software system that collects together a number of machine learning schemes and allows users to easily switch between schemes, using a common datafile standard and a unifying user interface to ease the testing and interpretation of results.

The machine learning research group at the University of Waikato has constructed this workbench to allow users to access a variety of machine learning techniques for the purposes of experimentation and comparison using real world data sets. The workbench is not a single program, but rather a set of tools bound together by a common user interface.

The WEKA workbench differs from other machine learning environments in that its target user is a domain expert, in this case an agricultural scientist, who wants to apply machine learning to real world data sets. Other systems are intended for use by machine learning researchers and programmers developing and evaluating machine learning schemes. The WEKA workbench concentrates on simplifying access to the schemes, so that their performance can be evaluated on their own.

WEKA currently includes a number of different machine learning schemes, summarized below. In a typical session, a user might select a data set, run several different learning schemes on it, exclude and include different sets of attributes, and make comparisons between the resulting concepts. Output from each scheme can be viewed in an appropriate form, for example as text, a tree or a graph.

	Scheme	Learning approach	Reference
Unsupervised	AUTOCLASS	Bayesian clustering	Cheeseman *et al*. (1988)
	CLASSWEB	Incremental conceptual clustering	Fisher *et al*. (1987), Gennari (1989)
Supervised	C4.5	Decision tree induction	Quinlan (1992)
	1R AND T2	Simple rules	
	INDUCT	Complex rules	Gaines (1991)
	IB1-4, PEBLS and K*	Instance based learning	
	M5’	regression model trees	
	FOIL	First-order inductive learner for relational rules	Quinlan (1990), Quinlan (1991), Quinlan *et al*. (1993), Cameron-Jones *et al*. (1993)

Machine learning schemes currently included in the WEKA workbench

Further information on the WEKA workbench, the activities of the machine learning group, and its publications can be obtained at http://www.cs.waikato.ac.nz/~ml.

1.3 Characteristics of the datasets

The table below gives an overview of the datasets tested. Instances refers to the number of observations, or "rows" in the dataset, while attributes refers to the number of variables, or "columns" recorded for each observation instance. In some cases, additional attributes were derived, or calculated from the original attributes in the provided dataset prior to machine learning runs, and these are indicated by an * in the table below. Attributes types are indicated as the following:
enumerated: data point is one of a set of discrete numbers or symbols, not necessarily in a continuous range. Numbers have no inherent meaning as a value other than symbolic. Comparison tests only on equality. The attribute value represented by the number 4 is not twice as big as the number 2. Examples: 1,4,9.2, high, a+, yes.
real: data point is in a continuous range of decimal numbers. Any number in the range is valid. Comparison tests include greater than, less than. Examples (for a range between -4.0 and 9.5): 4.1, 6.324, -2.5
integer: Non decimal number, usually positive. Comparison tests on greater than, less than. Numbers have order.

Section	Dataset Name	# of Instances	# of Attributes	enumerated	real	integer
2	Apple	1662	16	7	6	3
3	Bulls	90	8	3	5	
4	Eucalypt	736	20	6		
5	Grub Damage	155	9	5	3	1
6	Grub Rain	19	6	1	5	
7	Grower	22	15	1	14	
8	Pasture	36	25	2	15	8
9	Pea Seed	51	15	1	13	1
10	Slugs	100	9*	5	3	1
11	Squash	261	24	3	6	16
12	Valley	878	15	1	13	1
13	Venison	21,448	23	13	1	9
14	Wasp	506	13	6	5	2
15	White Clover	63*	32*	5	27	

1.4 Classes and accuracy

When a dataset is processed through the WEKA workbench, the objective is usually to partition the instances into "classes" of one of the attributes. For example, in trial number 8, pasture production, the class attribute was pasture production class, which could be one of three enumerated values, namely lo, med and hi. Machine learning then attempts to construct decision trees, based on the other attributes selected for the run, which will result in all instances being classified into those classes. Because the machine learning scheme tries to generate a "best fit" set of decision rules, those rules do not necessarily allow for all instances to be correctly put into their proper class, as specified in the attribute for that instance. To test itself, the machine learning scheme usually automatically calculates a measure of accuracy, that

is the percent of instances that are correctly placed in their proper class for each run processed through it.

Hypothetically, one could generate a set of rules, the number of which would equal the number of instances for the test dataset, which would classify each instance with 100% accuracy. Thus, for a 1000 instance dataset, a 1000 rule decision tree would be required. However, to better understand the phenomena being studied, we are willing to trade off 100% accuracy for smaller decision trees that give a clearer understanding of the relationship of the most important attributes in determining the proper class. The machine learning schemes thus allow for parameters to be specified which will determine the balance between trees with large numbers of rules and high accuracy, to much smaller and simpler rulesets with perhaps only slightly decreased accuracy. The datasets analysed in this study have been put through this iterative procedure to determine the best balance between accuracy and decision tree size.

1.5 WEKA Workbench functions and machine learning schemes used

The following WEKA Workbench functions and machine learning schemes were used. The attribute editor supports creation of new, or derived attributes based on mathematical or logical combinations of existing attributes.

Section	Dataset Name	Attribute editor	C4.5	1Rw	Autoclass
2	Apple	yes	yes		
3	Bulls		yes		
4	Eucalypt	yes	yes		
5	Grub Damage	yes	yes		
6	Grub Rain	yes	yes		
7	Grower		yes	yes	
8	Pasture	yes	yes		
9	Pea Seed	yes	yes	yes	
10	Slugs	yes	yes		
11	Squash		yes	yes	
12	Valley	yes	yes		yes
13	Venison	yes	yes	yes	
14	Wasp		yes	yes	
15	White Clover	yes	yes		

1.6 The results of the tests

The following sections contain a short summary of the tests undertaken with the datasets supplied. In each section, most of the following are given:

identification information
objectives of the original research
objective of the machine learning test
description of the size of the dataset
description of each attribute
class information

profiles of key attributes
machine learning processes undertaken
indicators of size of decision tree and errors
samples of the decision tree in graphical form
discussion of results
conclusions as to applicability of machine learning to this dataset

2. APPLE BRUISING

Data source	Brian De la Rue Lincoln Technology Hamilton
Date received	3 November, 1995
Assigned to	Colleen Burrows, Research programmer
Data location	/home/ml/datasets/Summer1995/apples-95-96/REPORT/original.arff /home/ml/datasets/Summer1995/apples-95-96/REPORT/results.arff

2.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To determine which attributes contribute to apple bruise sizes.

Summary of original research results
The original research results are currently unknown.

2.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Machine learning objective
To determine which attributes contribute highly to apple bruise sizes.

Dataset description
The dataset descriptions include instance, attribute, and class information.

The data is complete except 'BruiseWidthEquator' is present for only one impact energy value for each code. Instances in Code '95GS-A', '95GS-C', and '95GS-D', which are significantly smaller than instances for '95GS-B', should all have similar values for 'FruitRadius'.

Instance information
The original dataset contained attributes with large numbers of missing vlaues, these were removed for analyses purposes. The factors present are only a representation of all the features required to describe apple bruising.

Original dataset : 1662 instances.
Instances where 'Code' equals '95GS-E' and '95GS-F' were deleted from the dataset. Remaining instances totalled 1536.

Attribute information
Number of attributes : 16 (15 plus one predefined class).

The attributes are :

1. Code (bruise location - class value)
   type : enumerated
   values : 95GS-A, 95GS-B, 95-GS-C, 95GS-D
   missing : 0
2. HarvestTime
   type : enumerated
   values : mid
   missing : 0
3. ImpactSurface
   type : enumerated
   values : Steel, Fruit
   missing : 0
4. ImpactEnergy
   type : enumerated
   values : 0.019, 0.037, 0.056, 0.075, 0.093. 0.115
   missing : 0
5. FruitNumber
   type : enumerated
   values : 1 - 16
   missing : 0
6. BruiseLabel
   type : enumerated
   values : A, B, C, D
   missing : 0
7. FruitRadiusEquator
   type : integer
   values : n/a
   missing : 1536
8. FruitRadiusPolar
   type : integer
   values : n/a
   missing : 1536
9. FruitRadiusCombined
   type : integer
   values : 25.0 - 62.0
   missing : 0

10. ContactWidthEquator
    type : real
    values : 8.0 - 23.0
    missing : 0
11. ContactWidthPolar
    type : real
    values : 8.0 - 23.0
    missing : 0
12. BruiseWidthEquator
    type : real
    values : 10.0 - 20.0
    missing : 0
13. BruiseWidthPolar
    type : real
    values : 0.0 - 21.0
    missing : 0
14. BruiseDepthBottom
    type : real
    values : 0.0 - 7.0
    missing : 0
15. BruiseDepthTop
    type : real
    values : 0.0 - 5.0
    missing : 0
16. VisibleExternally
    type : enumerated
    values : 0 (no), 1 (yes)
    missing : 0

Class information

A number of classes were used throughout the analysis. These were :

'Code' - predefined class values

1. 95GS-A
2. 95GS-B
3. 95GS-C
4. 95GS-D

'CodeRadi2' - based on 'Radius' and 'Code' values

1. Alow
2. Amed
3. Ahi
4. Blow
5. Bmed
6. Bhi
7. Clow
8. Cmed
9. Chi
10. Dlow
11. Dmed
12. Dhi

*‘ContactArea’ - ‘ContactWidthEquator’ * ‘ContactWidthPolar’*

1. small
2. med
3. large
4. vlarge

‘BruiseDiameter’ - based on ‘BruiseWidthPolar’

1. SmallDiam
2. MediumDiam
3. LargeDiam
4. VeryLargeDiam

The distribution of the classes are :

‘Code’

1. 95GS-A
2. 95GS-B
3. 95GS-C
4. 95GS-D

‘CodeRadi2’

1. Alow
2. Amed
3. Ahi
4. Blow
5. Bmed
6. Bhi
7. Clow
8. Cmed
9. Chi
10. Dlow
11. Dmed
12. Dhi

‘ContactArea’

1. small
2. med
3. large
4. vlarge

‘BruiseDiameter’

1. SmallDiam
2. MediumDiam
3. LargeDiam
4. VeryLargeDiam

Data analyses procedure	Results
Three 'bruise size' classes were determined from a 'bruise area' calculation : • none = 0 • small < 1 • large > 1	
Experiments on the training set included : 1. C4.5 was run with 'Code' as the class and the attributes : 'ImpactEnergy' 'FruitRadiusCombined' 'ContactWidthEquator' 'ContactWidthPolar' 'BruiseDepthBottom' 'BruiseDepthTop' 'VisibleExternally'	1. The simplified decision tree was : size : 35(nodes) with 4.6% error dominant attributes : 'ImpactEnergy' 'BruiseDepthBottom' 'Code' This tree seems quite lopsided since the first node states that if the impact energy is greater than 0.019, then the bruise size will be large, and the other nodes include instances with impact energy at 0.019.
2. C4.5 was run with the attributes (above) except : 'BruiseDepthBottom' 'BruiseDepthTop'	2. The simplified decision tree had: size : 25 with 8.2% errors dominant attributes : 'ImpactEnergy' 'Code' 'VisibleExternally' Again the tree was lopsided.
3. C4.5 was run with the attributes: 'BruiseWidthPolar' 'BruiseDepthTop' 'BruiseDepthBottom' Each 'Code' group was divided into three categories - low, medium, high - based on the 'FruitRadius' value. This new attribute 'CodeRadi2' was used instead of 'Code' and 'FruitRadius'.	3. The simplified decision tree had: size : 117 with 37.8% errors This result may indicate that the three attributes are equally important in relationship to impact energy

4. C4.5 was run using 'sizebru2' with attributes : 'ImpactEnergy', 'ContactWidthEquator', 'ContactWidthPolar', 'BruiseDepthBottom', 'BruiseDepthTop' 'VisibleExternally' 'CodeRadi2'	4. The simplified decision tree had : *size : 5 with 5.7% errors* This resulted in : if the 'BruiseDepthBottom' is zero then there is no bruise, or if the 'ImpactEnergy' is 0.019 then the bruise is small, otherwise the bruise will be large. This indicates that the size of the bruise is wholly dependent on the impact energy.
A new attribute 'ContactArea' was created by multiplying values of 'ContactWidthEquator' by values of 'ContactWidthPolar'. Enumerated values were created with values : • small < 150 • med < 230 • large < 330 • vlarge > 330	
5. C4.5 was used with 'ContactArea' as a class with : 'ImpactEnergy' 'BruiseDepthBottom' 'BruiseDepthTop' 'BruiseWidthPolar' 'VisibleExternally' 'CodeRadi2'	5. The simplified decision tree had : size : 93 with 16% errors dominant attributes : 'ImpactEnergy' 'CodeRadi2' This indicates that impact energy had more effect on the bruise contact area than the bruise size.
A new attribute 'BruiseDiameter' was created from 'BruiseWidthPolar' where : • SmallDiam < 7 • MediumDiam < 13 • LargeDiam < 17 • VeryLargeDiam < 17	

6. C4.5 was run with 'BruiseDiameter' as the class with : 'ImpactEnergy' 'BruiseDepthBottom' 'BruiseDepthTop' 'VisibleExternally' 'ContactWidthEquator' 'ContactWidthTop' 'CodeRadi2'	6. The simplified decision tree had: size : 103 with 17.3% errors dominant attributes : 'ImpactEnergy' 'BruiseDepthBottom' 'ContactWidthEquator'
7. C4.5 was run using 'BruiseDiameter' as the class with the same attributes except 'Code and 'FruitRadiusCombined' were used instead of 'CodeRadi2'	7. The simplified decision tree had : size : 95 with 16.9% errors The result is almost identical to one above as 'Code' and 'FruitRadiusCombined' appear so low in the tree.

2.3 Discussion of results

Apples within a certain code were not of constant radius as was originally thought. 'ImpactEnergy' seems to contribute most to the size of the bruise and the contact area.

2.4 Conclusion

The results seem favourable to machine learning. One tree (size 5 with 5.7% errors) meets C4.5's requirements to distinguish a good tree. It has a small number of nodes, and has few errors. Contributing to this outcome is the fact that there are no missing values in the dataset. Consequently, noise will be reduced. Also, a large number of instances were analysed with a good number of attributes (not too many and not too smaller amount). Nearly half of the attributes contained enumerated value types. This often enables C4.5 to categorise instances more easily.

The ability to create new attributes enabled initial analyses techniques to be extended

3. BULLS

Data source	Jenny Jago Animal Behaviour and Welfare Research AgResearch, Ruakura Research Centre Hamilton
Report date	December 1995
Assigned to	Stuart Yeates, Research programmer
Data location	/home/ml/datasets/Summer1995/REPORT/original.arff /home/ml/datasets/Summer1995/REPORT/results.arff

3.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To find which treatments of castration give the heaviest bulls with the least violent behaviour.

Summary of original research results
The original research results are currently unknown.

3.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To find which treatments of castration give the heaviest bulls with the least violent behaviour, determine differences between different classes of bulls, and discover a relationship between testosterone level and class.

Dataset description
The dataset descriptions include instance, attribute, and class information. The dataset is very ‘wide’, with more attributes than instance.

Instance information
All instances contain some missing values and noise. Most instances have missing values for the first one or two attributes in the time series. There are no scrotal circumference or testes length present for group seven.

Original dataset : 90 instances.

Final dataset : 90 instances.

Attribute information

There were a number of attributes contained within the original dataset. These attributes can be grouped into eight categories. These categories are :

1. tag (unique identifier)
   type : enumerated
   values : 1 - 98
   missing : 0
2. group (treatment group)
   type : enumerated
   values : 1,2 immunised at 4, 4½, and 7½ months
   3,4 entire bulls
   5,6 immunised at 7½, and 8 months of age
   7 steer (castrated at 2 months of age)
   8,9 immunised at 2, 2½, 4, and 7½ months
   missing : 0
3. origin (farm of bull)
   type : enumerated
   values : Candy, McFarlane, Ferguson, Clarke
   missing : 0
4. legwear (riding behaviour of different dates)
   type : real
   values : 0.0 - 4.0
   missing : 1
5. scrotal (scrotal circumference at different dates)
   type : real
   values : 10.0 - 37.0
   missing : (on average) 10
6. testes (testes length on different dates)
   type : real
   values : 5.0 - 17.0
   missing : (on average) 10
7. testos (testosterone level on different dates)
   type : real
   values : 0.042 - 27.77
   missing : (on average) 2
8. wt (weight on different dates)
   type : real
   values : 58.5 - 429.0
   missing : (on average) 5

Class information

Two classes were used to classify the data. These were :

wt_final - based on last weight data point in attribute 'Weight'
1. (lightest) to 6. (heaviest)

group1
1. to 9.

The distribution of the classes are :
wt_final

1.	0
2.	23
3.	25
4.	19
5.	18
6.	5

group1
Each value for New_Treatment_Groups has 10 instances.

Analysis by 1Rw produced the following results :
wt_final
default class : 3 with 25/90 accuracy
best hypothesis : tag1 with 100% accuracy

group1
default class : 1 with 10/90 accuracy
best hypothesis : tag1 with 100% accuracy

Data analysis procedure	**Results**
1. C4.5 was used with 'group1' as the class with the testosterone level attributes.	1. The simplified decision tree was : *size : 23 (nodes) with 15% errors.*
2. C4.5 was run using 'wt_final' as the class with the testosterone level attributes.	2. The simplified decision tree was : size : 17 with 40% errors.
3. C4.5 was run using 'group1' as the class with the weight attributes.	3. The simplified decision tree was : size : 33 with 40% errors

3.3 Discussion of results

The resulting trees indicate strong links to testes length and testosterone which in turn are strongly linked to the final weight. The decision trees were overfitted in comparison to the data producing trees which correctly classify all of the instances in the training set but classify unseen examples relatively poorly.

3.4 Conclusion

The dataset was not very suitable to machine learning due to the time series nature of the data.

Follow up notes
Jenny Jago commented that machine learning did not find as many relationships as were expected.

4. EUCALYPTUS SOIL CONSERVATION

Data source Bruce Bulloch
128 Cook Street
Palmerston North

Report date January 1996

Assigned to Kirsten Thomson, Research programmer

Data location /home/ml/datasets/Summer1995/eucalyptus/REPORT/original.arff
/home/ml/datasets/Summer1995/eucalyptus/REPORT/resultsarff

4.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives

To determine which seedlots in a species are best for soil conservation in seasonally dry hill country. Determination is found by measurement of height, diameter by height, survival, and other contributing factors.

Summary of original research results

On the basis of the collected data (1991) larger growing species seem more suited to dry hill country areas, such as Pakaraka and Kahuiti. Domain knowledge allowed the researcher to determine that the larger trees would serve best in conservation/production/agroforestry regimes, the smaller at wider spacings in conservation/pastoralism regimes. A further note to the study, stated that provenance comparisons within species generally support the findings of more extensive genetic improvement trials focused on the eucalypts with most potential for timber production.

4.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Machine learning objective

To contrast seedlots and determine if one within a species is better than another (within the experimental guidelines), ie. dry hill country. This is to be achieved by contrasting relationships and determining which attributes contribute most to a good or better result.

Dataset description
The dataset descriptions include instance, attribute, and class information.

The original data was contained in two datasets. The first describing the site and location details, the other eucalyptus growth, in accordance to the different areas. Attributes one - 10 contain the species details. Attributes 11 - 21 contain the site and location details.

It is important to note that eucalypt trial methods have changed over time; earlier trials included mostly 15 - 30cm tall seedlings grown in peat plots and the later trials have included mostly three replications of eight trees grown. This change may contribute to less significant results.

Experimental data recording procedures which require noting include :

- instances with no data recorded due to experimental recording procedures require that the absence of a species from one replicate at a site was treated as a missing value, but if absent from two or more replicates at a site the species was excluded from the site's analyses.
- missing data for survival, vigour, insect resistance, stem form, crown form and utility especially for the data recorded from Morea Station; this could indicate the death of species in these areas or a lack in collection of data.

Instance information
A number of instances contained missing values and noise. The missing values usually represent a plant dying during the time of the experiment. Unknown values are represented by a '?'. The factors present are only a representation of all the features required to describe eucalyptus planting.

Original dataset : 736 instances.
Final dataset : 736 instances.

Attribute information
Number of attributes : 20.

The attributes are :

1. Sp (species)
   type : enumerated
   values : co, fr,.....,te
   missing : 0
2. PMCNo (seedlot number)
   type : integer
   values : 1.0 - 3,275.0
   missing : 7
3. DBH (best diameter base height, cm)
   type : real
   values : 0.58 - 42,085.0
   missing : 1

4. Ht (m)
   type : real
   values : 1.12 - 21.79
   missing : 0
5. Surv (survival)
   type : integer
   values : 1.0 - 100.0
   missing : 94
6. Vig (vigour)
   type : real
   values : 0.5 - 5.0
   missing : 69
7. Ins_res (insect resistance)
   type : real
   values : 0.0 - 4.5
   missing : 69
8. Stem Fm (stem form)
   type : real
   values : 0.0 - 5.0
   missing : 0
9. Crown Fm (crown form)
   type : real
   values : 0.0 - 0.5
   missing : 0
10. Brnch Fm (branch form)
    type : real
    values : 0.0 - 5.0
    missing : 69
11. Utility (utility rating)
    type : real
    values : 0.0 - 5.0
    missing : 94
12. Name (name of site;)
    type : enumerated (15)
    values : Clydebank, Craggy Range Road,.....,Wimbledon Species
    missing : 0
13. Abbreviation (site abbreviation)
    type : enumerated (15)
    values : Cly, Cra,.....,WSp
    missing : 0
14. Locality (where the site is in the North Island)
    type : enumerated
    values : Central_Hawkes_Bay, Northern_Hawkes_Bay,....,
    South_Wairarapa
    missing : 0
15. Map Ref (map location in the North Island, NZMS1)
    type : enumerated
    values : N135_382/137, N116_848/985,.....,N151_922/226
    missing : 0

16. Latitude approximation (South - deg.,min)
    type : enumerated
    values : 29_38, 39_00,.....,82_32
    missing : 0
17. Altitude (m)
    type : integer
    values : 70 - 300
    missing : 0
18. Rainfall (mm pa)
    type : integer
    values : 850 - 1,750
    missing : 0
19. Frosts (deg.C)
    type : integer
    values : -3.0 - -2.0
    missing : 0
20. Year (year of planting)
    type : integer
    values : 1980 - 1986
    missing : 0

Class information

There are two class values created for this dataset from existing attributes. These were:

New_Utility - based on existing utility values

1. none
2. low
3. average
4. good
5. best

*Bulk - DBH * DBH * Ht*

Enumerated, with many values.

The distribution of the classes are :

New_Utility

1.	none	180
2.	low	107
3.	average	130
4.	good	214
5.	best	105

Bulk

There is one class value for every instance.

Analysis by 1Rw produced the following results :

New_Utility

default class : good with 214/736

best hypothesis : ‘Vig’ with 50.8152% accuracy

Data analysis procedure	Results
The two data sets were combined to create on dataset with attributes from both.	
The original utility measurement was categorised into a new utility rating : • none < 0.1 • low < 1.48 • average < 2.65 • good < 3.83 • best < 5.00	
A number of tests were run on the data : 1. C4.5 was used with 'New_Utility' as the class with the attributes : 'Species' 'PMCNo'	1. The simplified tree had : size : 62 (nodes) with 47.7% errors dominant attributes : 'Species' Species divided the data more than the seed lot number, since many of species individuals were all in one class.
2. C4.5 was used with 'New_Utility' as the class with the attributes : 'Species' 'PMCNo' 'DBH' 'Ht' 'Surv' 'Vig' 'Ins_res' 'Stem_Fm' 'Crown_Fm' 'Branch_Fm'	2. The simplified tree had : size : 234 with 14.7% errors dominant attributes are 'Vigour', 'Species' and 'PMCNo'.
3. C4.5 was used with 'New_Utility' as the class with the attributes with the attributes above including : 'Rep' 'Locality' 'Map_Ref' 'Latitude' 'Altitude' 'Rainfall' 'Frosts'	3. The simplified tree had : size : 227 with 7.7% errors. dominant attributes are 'Year' dividing the data at 1983, and 'Vigour', 'Survival' and 'Height'

A new attribute 'Bulk' was created based on the 'DBH'*'DBH'*'Ht' attributes. This created may values, generally, each instance had it's own class value.	
4. C4.5 was used with 'Bulk' as the class with the attributes with all other attributes except : 'Utility' 'DBH' 'Height'.	4. The simplified tree has : size : 151 with 26.5% errors dominant attributes are 'Year' at 1983, and 'Vigour' and 'Species'.
The 'Species' attribute was altered. The values 'ma' were changed to 'mn' (there was an error in the data). The instances for Morea Station and Wensleydale were deleted from the dataset (they had no survival rates, and were irrelevant to the outcome). The 'PMCNo' was changed to an enumerated value.	The last three decision trees used vigour as one of the first dividers of the data. In all trees, the data was divided firstly at vigour 2.3 and then at vigour 1.3 where those instances with vigour less than or equal to 1.3 had a utility rating of 'none'. There are missing data values for survival, vigour, insect resistance, stem form, crown form, branch form and utility especially from data from Morea Station which may indicate death of species in these areas or a lack in collection of data. No data was recorded for 1984 and 1985 and very little data was recorded for 1986 which explains the first division in two of the decision trees at year 1983 where those greater than 1983 give a utility rating of 'none'.
5. C4.5 was used with 'New_Utility' as the class with : 'PMCNo' (enumerated) 'Species'	5. The simplified tree has : size : 94 with 50.1% errors dominant attributes are 'PMCNo'
6. C4.5 was used with 'New_Utility' as the class with : 'PMCNo' 'Species' 'Site'	6. The simplified tree has : size : 266 with 42.9% errors dominant attributes are 'PMCNo' then 'Abbrev'

7. C4.5 was used with 'New_Utility' as the class with : 'PMCNo' 'Species' 'DBH' 'Height' 'Surv' 'Vigour' 'Ins_res' 'Stem_Fm' 'Crown_Fm' 'Branch_Fm'	7. The simplified tree has : size : 126 with 18.9% errors dominant attributes are 'Vig', 'Surv' and 'PMCNo'.
8. C4.5 was used with 'New_Utility' as the class with the attributes above excluding 'Species'.	8. The simplified tree has : size : 128 with 31.8% errors dominant attributes are 'Vig', 'Surv' and 'PMCNo'.
9. C4.5 was used with 'New_Utility' as the class with the attributes above including : 'Rep' 'Locality' 'Map_ref' 'Latitude' 'Altitude' 'Rainfall' 'Frosts' 'Year'	9. The simplified tree has : size : 141 with 31.5% errors dominant attributes are 'Vig', 'Surv' and 'PMCNo'.
Instances with a 'New_Utility' rating of 'good' and 'best' were created into a new dataset.	
10. C4.5 was used with 'New_Utility' as the class on this new dataset with all of the attributes used above in 9.	10. The simplified tree has : *size : 17 with 20.7% errors* dominant attributes are 'Vig' and 'Crown_Fm' .

11. C4.5 was used with 'New_Utility' as the class with on this new dataset with : 'Replicate' 'Locality' 'Map_ref' 'Latitude' 'Altitude' 'Rainfall' 'Frosts' 'Year'	11. The simplified tree has : size : 10 with 27.9% errors dominant attributes are 'Locality' and 'Rep'.
The instances with 'none' and 'low' have been amalgamated into a 'lower' value. Those above these values have stayed the same.	
12. C4.5 was used with 'New_Utility' as the class with the attributes above including : 'PMCNo' 'Species' 'DBH' 'Height' 'Surv' 'Vigour' 'Ins_res' 'Stem_Fm' 'Crown_Fm' 'Branch_Fm'	12. The simplified tree has : size : 139 with 20.8% errors dominant attributes are 'Vig' and 'PMCNo'.
13. C4.5 was used with 'New_Utility' as the class with: 'Replicate' 'Locality' 'Map_ref' 'Latitude' 'Altitude' 'Rainfall' 'Frosts' 'Year'	13. The simplified tree has : size : 29 with 51.7% errors dominant attributes are 'Latitude'.

The results are presented at the end of this analyses.

4.3 Discussion of results

Due to initial analysis performed with 'PMCNo' as an integer (this was incorrect, and can be equated to comparing two colours ie. is one colour less than another), the first four results have been discarded. The proper type of 'PMCNo' is as an enumeration. Therefore, the results from five to 13 are more correct. From these the best results are where :

1. size : 126 with 18.9% errors (no. 7)
   for attributes 'PMCNo', Species' and site details
2. size : 17 with 20.7% errors (no. 10)
   for all tree details, but on the smaller dataset of only good and best results for 'New_Utility'
3. size : 139 with 20.8% errors (no. 12)
   for all attributes, but none and low outcome have been amalgamated into one category.

4.4 Conclusion

Ideally, trees and rules found from C4.5 analysis should be smaller and more accurate, for example, less than 20 nodes and 10 percent errors. The best decision trees, listed above, have error rates above 10 percent. The first has the lowest error rate found, the second and last have higher error rates. These error rates may be due to the fact that data was missing from a number of attributes. Some of these missing values may be unknown or unmeasured values, others may be missing (or unable to be measured).

Machine learning analyses looks for relationships or patterns in the data. Some data, may be different from an established pattern (found from other data within the dataset). Though the data may be correct (determined by domain knowledge), the differentiation may not be recognised by the algorithm used, consequently a higher error count is attributed to. In this case domain knowledge and researcher input is important to determine if the error count is feasible.

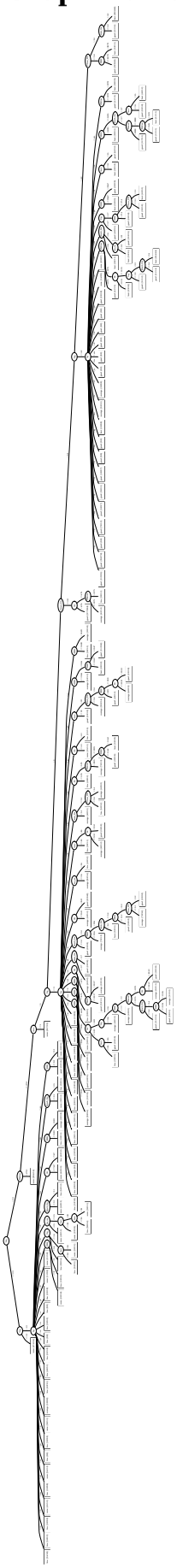
The large number of nodes in most of the trees can be attributed by the large number of attributes and large number of values within each of these attributes. The tree with the smallest nodes (size 17) was found from the smallest dataset with a smaller range of attributes (tree information only).

The results gained from machine learning seem interesting. Because of the higher number of errors, domain knowledge is required for further analysis of the results. The trees show which attributes and values contribute to good eucalyptus trees for soil conservation.

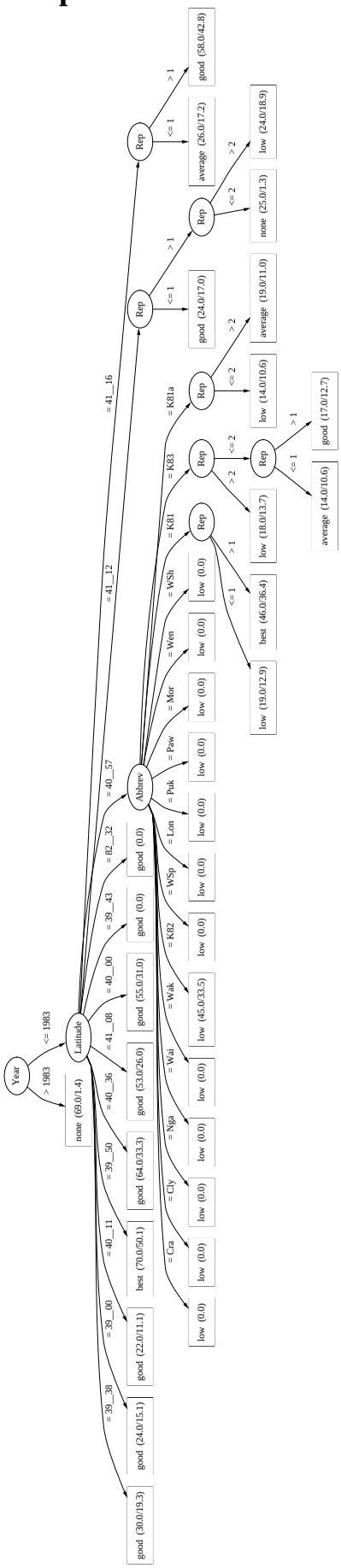
Follow up notes
Bruce Bulloch was very enthusiastic about the results and would like to work some more with machine learning on his dataset.

Reference : Bulluch B. T., (1992) Eucalyptus Species Selection for Soil Conservation in Seasonally Dry Hill Country - Twelfth Year Assessment New Zealand Journal of Forestry Science 21(1): 10 - 31 (1991)

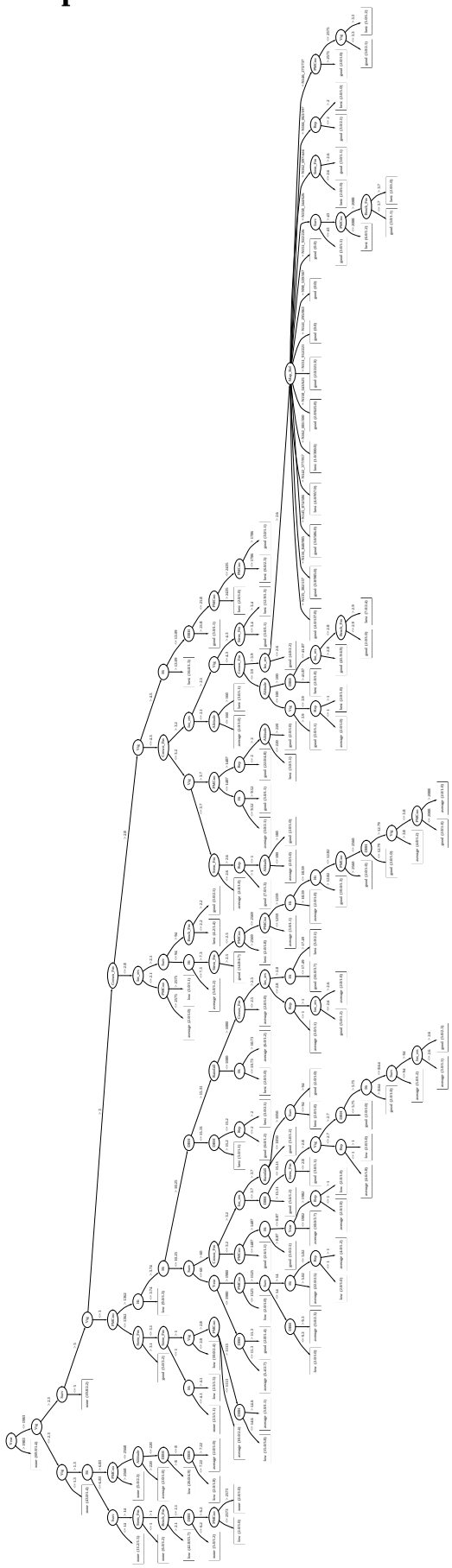
Graph 1 : Result 7



Graph 2 : Result 10



Graph 3 : Result 12



5. GRASS GRUBS AND DAMAGE RANKING

Data source R. J. Townsend
AgResearch
Lincoln

Date received January, 1996

Assigned to Colleen Burrows, Research programmer

Data location /home/ml/datasets/Summer1995/grass_grub/REPORT/ggdr_orig.arff
/home/ml/datasets/Summer1995/grass_grub/REPORT/ggdr.arff

5.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives

Grass grubs are one of the major insect pests of pasture in Canterbury and can cause severe pasture damage and economic loss. Pastoral damage may occur periodically over wide ranging areas. Grass grub populations are often influenced by biotic factors (diseases) and farming practices (such as irrigation and heavy rolling). The objective of the report was to report on grass grub population and damage levels to provide objective estimates of the annual losses caused by grass grubs.

Summary of original research results

Two studies were performed. The first from 1974 - 1985 which estimated the mean population density of grass grubs in the month of August ranged from 42 - 109/m^2. The highest densities were recorded for 1971 - 1981. The second study, 1986 - 1992 reported that the mean population density in September ranged from 37 - 114/m^2, with the highest density occurring in 1989. Grass grub population density in August/September was compared with weather patterns and a significant correlation was found with rainfall the preceding November. Average pasture damage over the complete period (1974 - 1992) was estimated as 8%. Peaks of 14, 20 and 12 % damage occurred in 1976, 1977 and 1981 in the first study and 15 and 18% damage peaks occurred in 1989 and 1990 during the second study period. Significant relationships were found between damage and grass grub density and between damage and autumn rainfall.

5.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To find a relationship between grass grub number, irrigation, and damage ranking for the period between 1986 and 1992.

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
A few instances, especially in ‘damage_rankRJT’ and ‘dry_or_irr’, contained missing values and noise. The factors present are only a representation of all the features required to describe damage caused by grass grubs.

Original dataset : 173 instances.
Final dataset : 155 instances.

Attribute information
Number of attributes : nine.

The attributes are :

1. Year_Zone
   type : enumerated (20 values)
   values : 0c, 0f, 0m,.....9c, 9f, 9m
   missing : 0
2. Year
   type : enumerated
   values : 1986-1992
   missing : 0
3. strip
   type : integer
   values : 1.0 - 10.0
   missing : 0
4. pk (paddock)
   type : integer
   values : 0.0 - 5.0
   missing : 0
5. damage_rankRJT (R. J. Townsend’s ranking)
   type : integer
   values : 0.0 - 5.0
   missing : 17
6. damage_rankALL
   type : enumerated
   values : 0.0 - 5.0
   missing : 5
7. GG_per_m2 (grass grubs per metre square)
   type : real
   values : 0.0 - 708.0
   missing : 12
8. dry_or_irr
   type : enumerated
   values : D (dryland), O (irrigated overhead), B (irrigated border dyke)

missing : 6

9. zone
   type : enumerated
   values : F (foothills), M (midplain), C (coastal)
   missing : 0

Class information

There has been one class created for this dataset.

GG_new

1. low
2. average
3. high
4. veryhigh

The distribution for the classes are :

1.	low	49
2.	average	41
3.	high	46
4.	veryhigh	19

Data analysis procedure	**Results**
The new class, GG_new, was determined from the attribute GG_per_m2, calculating : • low < 60 • average < 130 • high < 280; • veryhigh > 280 In all of the following C4.5 runs, the attributes 'year', 'strip', and 'paddock' were not used. They were unable to predict any other instance than those within the training set.	
Tests on the training set were : 1. C4.5 was used with the class GG_new and attributes: 'damage_rankRJT' 'damage_rankALL ' 'dry_or_irr' 'zone'	1. The simplified (pruned) decision tree was : size : 15(nodes) with 43.2% errors dominant attributes : 'damage_rankRJT' 'damage_rankALL'
2. C4.5 was run with GG_new as the class and attributes : 'dry_or_irr' 'zone'	2. The simplified decision tree was : size : 7 with 54.8% errors. dominant attributes : 'zone' 'dry_or_irr'

The data was split so that those irrigated (O or B) are classified as average, then of the dry plots, the foothills have low grass grubs and the coastal have high grass grubs.	
3. C4.5 was then run with 'damage_rankRJT' as the class and attributes : 'damage_rankALL' 'dry_or_irr' 'zone' 'GG_new'	3. The simplified decision tree was : size : 21 with 39.4% errors dominant attributes : 'damage_rankALL' 'GG_new' 'dry_or_irr'.
4. C4.5 was then run with 'damage_rankRJT' as the class and attributes : 'GG_per_m2' 'dry_or_irr' 'zone'	4. The simplified decision tree was size : 60 with 43.2% errors dominant attributes : 'GG_per_m2' 'dry_or_irr' 'zone'
5. C4.5 was run with 'damage_rankALL' as the class with attributes : 'damage_rankRJT' 'dry_or_irr' 'zone' 'GG_new'	5. The simplified decision tree was *size : 10 with 40.6% errors.* dominant attributes : 'damage_rankRJT' 'GG_new'
6. C4.5 was run with 'damage_rankALL' as the class with attributes : 'GG_per_m2' 'dry_or_irr' 'zone'	6. The simplified decision tree was size : 59 with 43.2% errors dominant attributes : 'GG_per_m2'

The three best results are presented at the end of this analyses.

5.3 Discussion of results

Small decision trees were created that show that the damage ranking of both RJT and ALL were related to the number of grass grubs. The zones and irrigation methods did not seem to play a consistent role in determining either the number of grass grubs when used with the damage rank attributes. However, when GG_new was used as the class with only 'zone' and 'dry_or_irr', C4.5 was able to separate the irrigated paddocks from the dry paddocks. When either of the damage rank attributes were used as a class, the other damage rank attribute was the most contributing one to the tree. This demonstrates that the rankings are similar. When

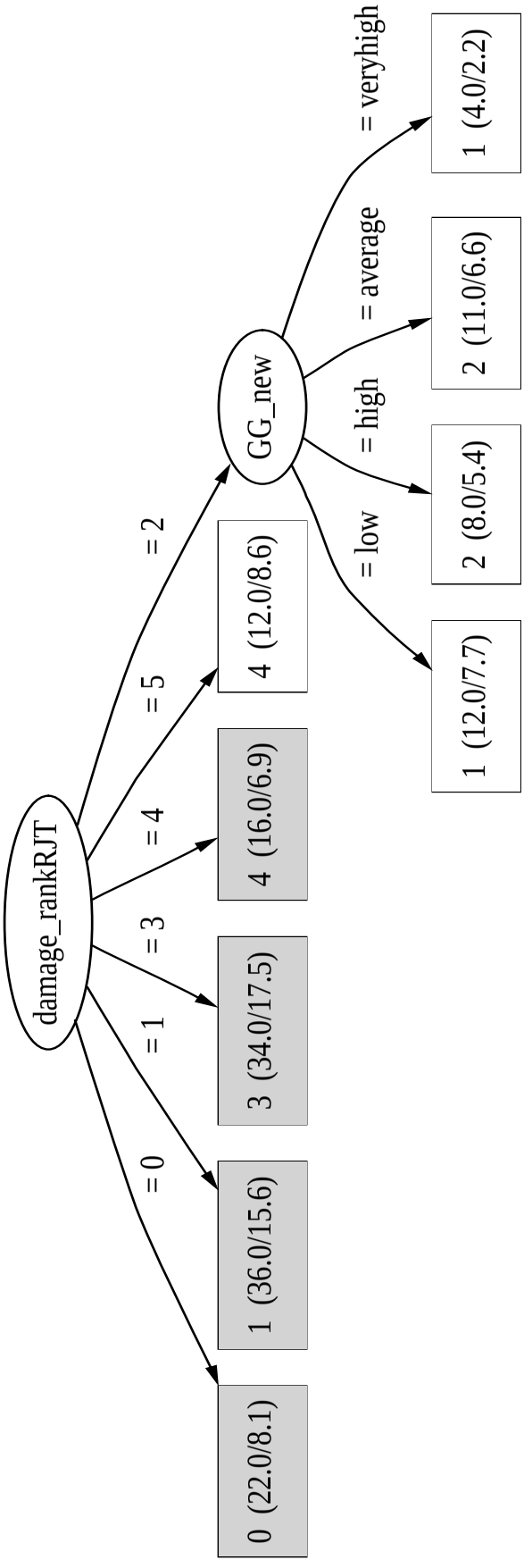
using one of the damage rank attributes as the class without the other damage rank attributes present, the tree is very large with the number of grass grubs playing the most significant role in determining the rank. In all decision trees, there are a large number of errors which questions the validity of the trees.

5.4 Conclusion

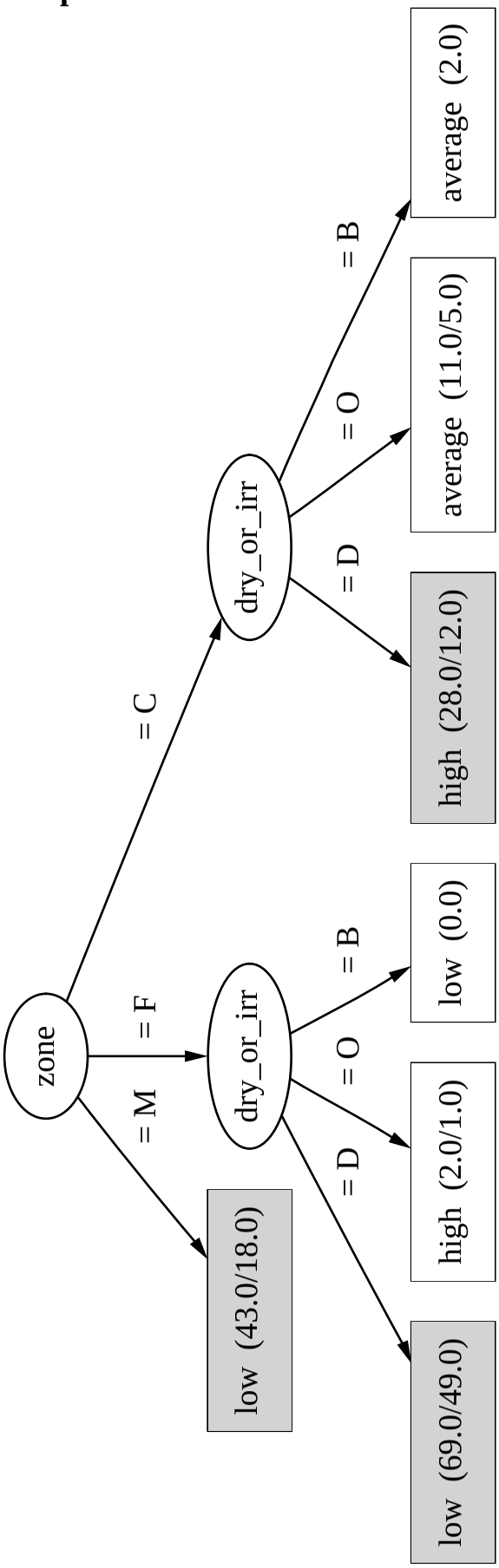
On the outset this dataset seemed to be quite favourable to machine learning since it has attributes which can be easily used as a class without any discretisation. However, even though the decision trees that were created were generally quite small they had a high number of errors. The high error counts are more likely to be due to the number of missing values for 'damagerankRJT' and 'damagerankALL'. The smaller trees can be accounted for by the number of attributes used in the dataset. In general, the smaller number of attributes produce a smaller tree (depending on the information contained within each dataset). Smaller error counts may have been found if all the types were enumerated into groups of between five to seven values.

Reference : Townsend R. J., Jackson T. A., Pearson J. F., French R. A. (1993) *Grass Grub Population Fluctuations and Damage in Canterbury*, Proc. 6th Australasian Grasslands Invert. Ecol. Conf. 1993, R. A. Prestidge, Ed.)

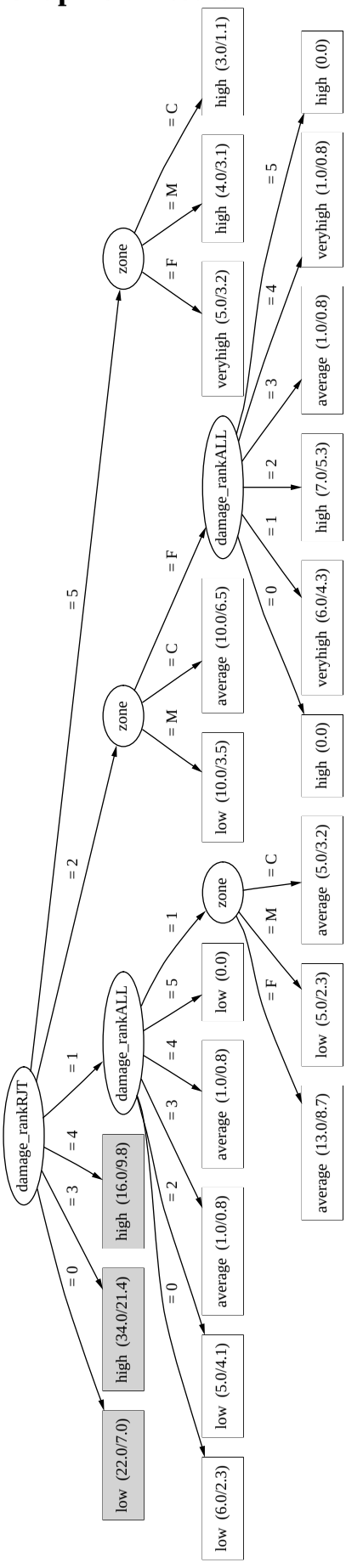
Graph 1 : Result 5



Graph 2 : Result 2



Graph 3 : Result 1



6. GRASS GRUBS AND RAINFALL

Data source David Saville
AgResearch
Lincoln

Date received December 1995

Assigned to Colleen Burrows, Research programmer

Data location /home/ml/datasets/Summer1995/grass_grub/REPORT/original.arff
/home/ml/datasets/Summer1995/grass_grub/REPORT/ggr.arff

6.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives

Grass grubs are one of the major insect pests of pasture in Canterbury and can cause severe pasture damage and economic loss. Pastoral damage may occur periodically over wide ranging areas. Grass grub populations are often influenced by biotic factors (diseases) and farming practices (such as irrigation and heavy rolling). The objective of the report was to report on grass grub population and damage levels to provide objective estimates of the annual losses caused by grass grubs.

Summary of original research results

Two studies were performed. The first from 1974 - 1985 which estimated the mean population density of grass grubs in the month of August ranged from 42 - 109/m^2. The highest densities were recorded for 1971 - 1981. The second study, 1986 - 1992 reported that the mean population density in September ranged from 37 - 114/m^2, with the highest density occurring in 1989. Grass grub population density in August/September was compared with weather patterns and a significant correlation was found with rainfall the preceding November. Average pasture damage over the complete period (1974 - 1992) was estimated as 8%. Peaks of 14, 20 and 12 % damage occurred in 1976, 1977 and 1981 in the first study and 15 and 18% damage peaks occurred in 1989 and 1990 during the second study period. Significant relationships were found between damage and grass grub density and between damage and autumn rainfall.

6.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective

To find relationships between the number of grass grubs present and the amount of rain that occurred in November, Summer, and Autumn or the percentage damage to an area and the amount of rain that occurred in November, Summer, and Autumn.

Dataset description

The dataset descriptions include instance, attribute, and class information.

Instance information

There were no missing values. The factors present are only a representation of all the features required to describe grass grubs in relation to rain or percentage damage in relation to rain in the months of November, Summer, and Autumn.

Original dataset : 19 instances.
Final dataset : 19 instances.

Attribute information

Number of attributes : six.

The attributes are :

1. year
   type : enumerated
   values : one value for every instance represented by year
   missing : 0
2. gg_per_m2 (grass grubs per metre square)
   type : real
   values : 37.0 - 114.0
   missing : 0
3. pct_damage (percent damage)
   type : real
   values : 2.8 - 19.9
   missing : 0
4. nov_rain
   type : real
   values : 12.0 - 119.0
   missing : 0
5. summer_rain
   type : real
   values : 98.0 - 308.0
   missing : 0
6. autumn_rain
   type : real
   values : 107.0 - 365.0
   missing : 0

Class information

Two class values were created for this dataset

'gg_new'

1. low
2. below_avg
3. above_avg
4. high

'pct_dam_new'

1. low
2. below_avg
3. above_avg
4. high

The distribution of the classes are :

'gg_new'

1.	low	4
2.	below_avg	4
3.	above_avg	5
4.	high	6

'pct_dam_new'

1.	low	5
2.	below_avg	6
3.	above_avg	3
4.	high	5

Analysis with 1Rw produced the following results :

'gg_new'

default class : 'high' with 6/19

best hypotheis : 'Year' with 100% accuracy.

'pct_dam_new'

default class : 'below_avg' with 6/19

best hypothesis : 'Year' with 100% accuracy.

Data analysis procedure	**Results**
Two new enumerated attributes were made 'gg_new' was created from 'gg_per_m2' with : • low < 50 • below_avg < 75 • above_avg < 90 • high > 90	

'pct_dam_new' was created from 'pct_damage' with : • low < 5.5 • below_avg < 7.2 • above_avg < 10 • high > 10	
1. C4.5 was run with 'gg_new' as the class and attributes 'pct_damage' 'nov_rain' 'summer_rain' 'autumn_rain'.	1. The simplified decision tree was size 7 (nodes) 26.3% errors dominant attributes were 'autumn_rain' (dividing at 165) and 'pct_damage' (dividing at 8.9 and at 5.2).
2. C4.5 was run with 'gg_new' as the class and attributes : 'nov_rain' 'summer_rain' 'autumn_rain'.	2. The simplified tree had size : 11 26.3% errors dominant attribute : 'autumn_rain'
3. C4.5 was run with 'pct_dam_new' as the class and attributes : 'gg_per_m2' 'nov_rain', 'summer_rain' 'autumn_rain'.	3. The simplified decision tree was *size : 11 21.15% errors* dominant attributes : 'gg_per_m2' 'nov_rain 'autumn_rain'
4. C4.5 was run with 'pct_dam_new' as the class and attributes : 'nov_rain', 'summer_rain' 'autumn_rain'.	4. The simplified decision tree was size : 9 26.3% errors dominant attribute : 'autumn_rain'

The three best results are presented at the end of this analyses.

6.3 Discussion of results

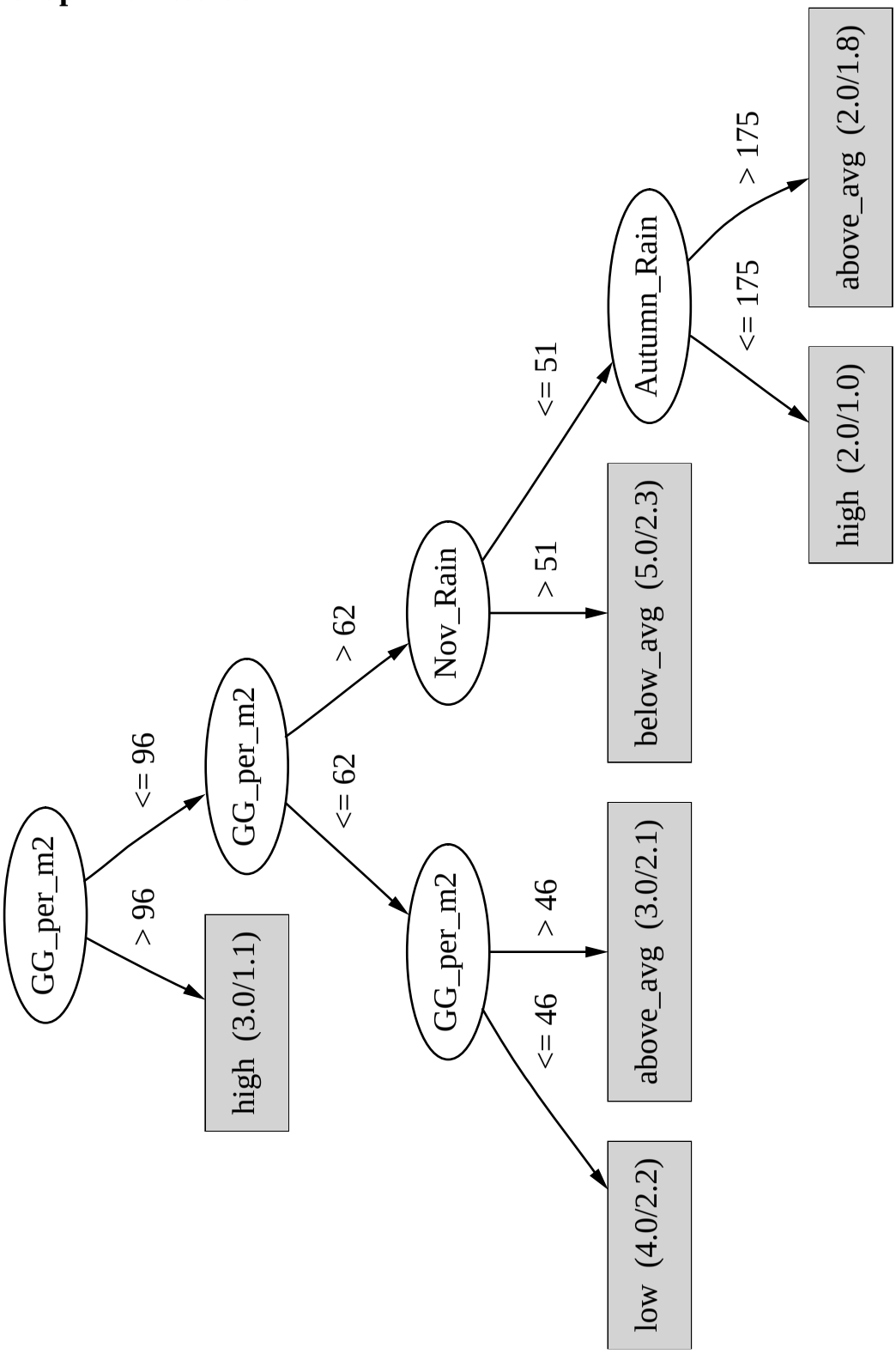
Grass grubs per metre square and percent damage are related as each appears in the tree of the other at a high position. The trees are not very accurate (21.1 and 26.3% errors) which is probably due to the small number of instances available.

6.4 Conclusion

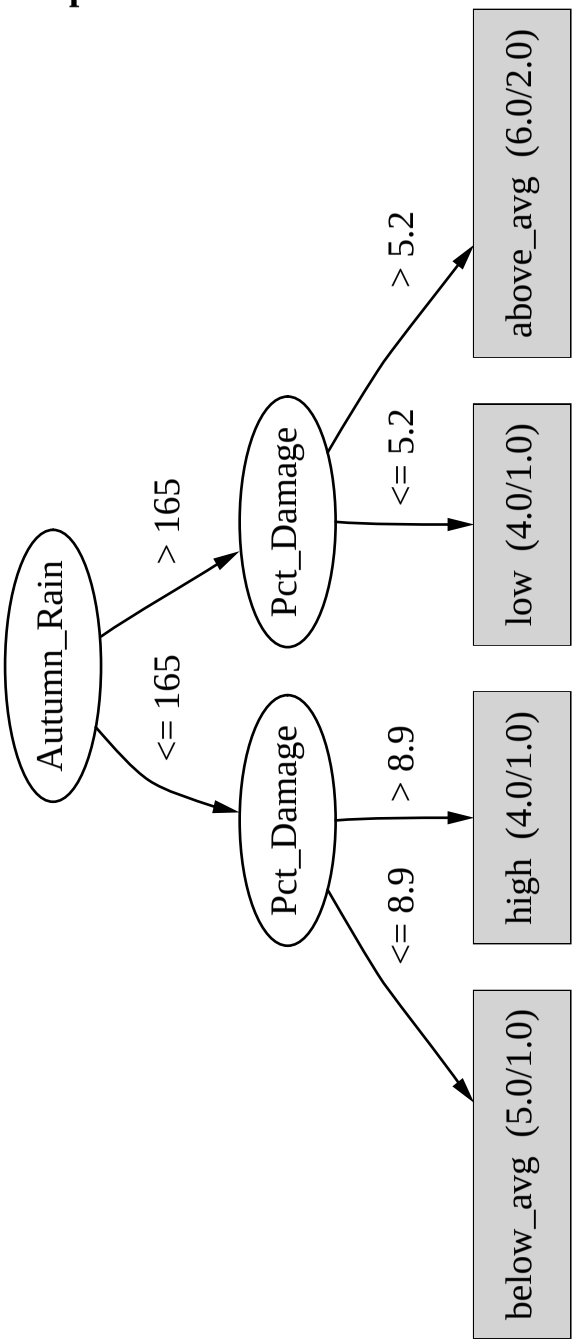
The errors remained consistent for three decision trees which may indicate that splitting (error levels) occurs only on one attribute. Presenting the 'rain' values as enumerated types between five and seven values may have achieved fewer errors. No data was missing so the dataset made a good training set. However, the small number of instances in the dataset produce results which are too general. A larger dataset would make the results more feasible. The small number of attributes probably contributed to the small decision trees which were easy to read.

Reference : Townsend R. J., Jackson T. A., Pearson J. F., French R. A. (1993) *Grass Grub Population Fluctuations and Damage in Canterbury*, Proc. 6th Australasian Grasslands Invert. Ecol. Conf. 1993, R. A. Prestidge, Ed.)

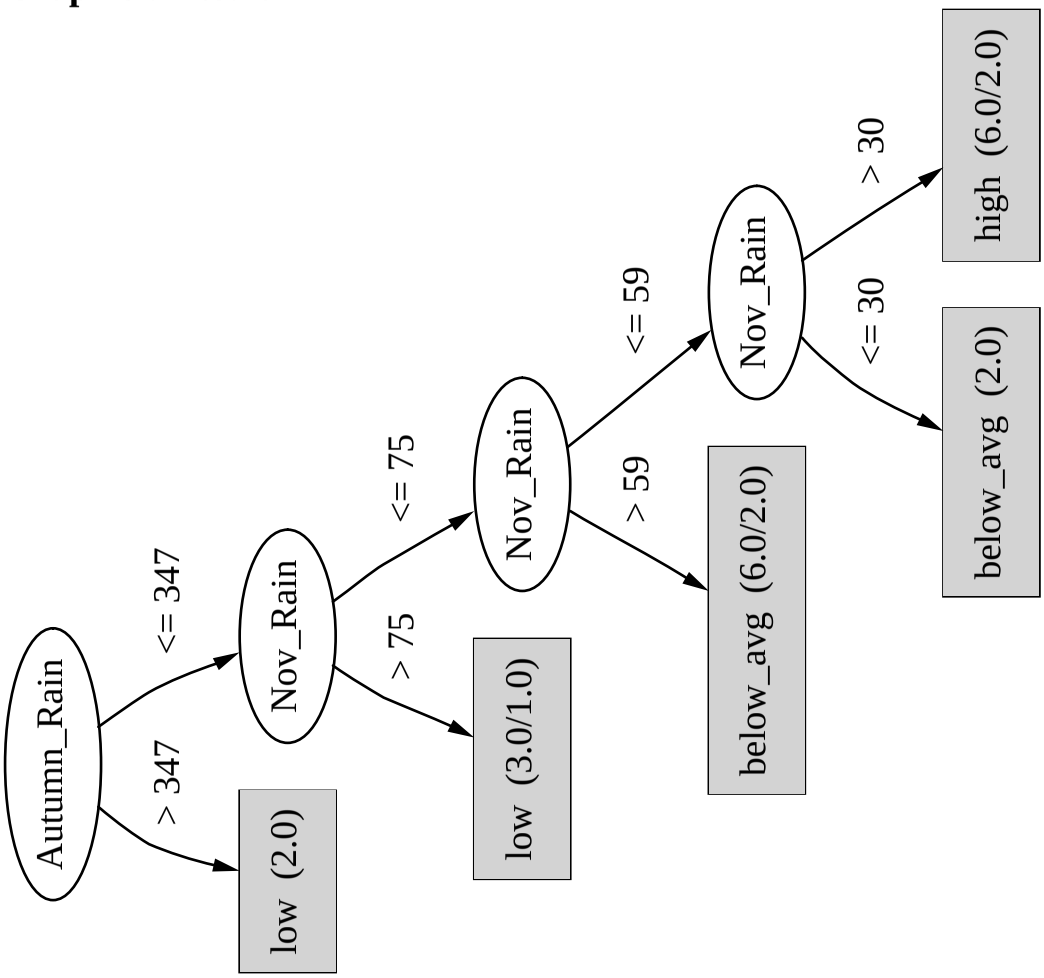
Graph 1 : Result 3



Graph 2 : Result 1



Graph 3 : Result 4



7. GROWER QUESTIONNAIRE

Data source Matthew Laurenson
Hort Research
Batchelor Research Centre
Palmerston North

Report date January 1996

Assigned to Colleen Burrows, Research programmer

Data location /home/ml/datasets/Summer1995/survey/spotchec.arff

7.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
The original objective differs from the machine learning objective quite significantly. The original objective was to collect data about grower practices and weather conditions with regard to fruit size and quality and to predict from this data the software buying intentions of either growers or consultants in this field.

The class attribute should indicate the person filling out survey (grower or consultant) and the other attributes should describe how much money, out of $100, would be allotted for software concerning the various issues implied by the attribute name.

The data was collected via a survey.

Summary of original research results
The original research results are currently unknown.

7.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To find some relationships which would indicate that the respondent was either a grower or a consultant depending on the answers given in the survey.

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
None of the instances contained missing values. The factors present are only a representation of all the features required to describe software buying intentions of either growers or consultants.

Original dataset : 22 instances.
Final dataset : 22 instances.

Attribute information
Number of attributes : 15 (14 actual attributes plus one predefined class).

The attributes are :

1. Class
   type : enumerated
   values : G (grower), C (consumer)
   missing : 0
2. Ascospore
   type : integer
   values : 0.0 - 30.0
   missing : 0
3. Expert
   type : integer
   values : 0.0 - 30.0
   missing : 0
4. Additional_Elements
   type : integer
   values : 0.0 - 50.0
   missing : 0
5. Replanting
   type : integer
   values : 0.0 - 0.0
   missing : 0
6. Fireblight
   type : integer
   values : 0.0 - 25.0
   missing : 0
7. ERM
   type : integer
   values : 0.0 - 35.0
   missing : 0
8. Chemical_Thinning
   type : integer
   values : 0.0 - 40.0
   missing : 0
9. Fruit_Size
   type : integer
   values : 0.0 - 25.0
   missing : 0

10. GDD_Calc
    type : integer
    values : 0.0 - 22.0
    missing : 0
11. Forecast
    type : integer
    values : 0.0 - 100.0
    missing : 0
12. Weather_data
    type : integer
    values : 0.0 - 45.0
    missing : 0
13. Frost
    type : integer
    values : 0.0 - 18.0
    missing : 0
14. Nutrition
    type : integer
    values : 0.0 - 25.0
    missing : 0
15. Safety
    type : integer
    values : 0.0 - 12.0
    missing : 0

Class information

For both datasets there are three class values.

1. G (grower)
2. C (consumer)

The distribution of the classes are :

1. G (grower) 16
2. C (consumer) 6

1Rw analysis on the dataset produced the following results :

default class : 'G' with 16/22
best hypothesis : 'ERM' with 81% accuracy.

Data analysis procedure	**Results**
1. C4.5 was run using the 'class' attribute as the class with all other attributes.	1. The simplified decision tree was : *size : 9 (nodes) with 4.5% (error)* dominant attributes :c 'Ascospore' 'Chemical_Thinning' 'Weather_data' 'GDD_Calc'.

The results are presented at the end of this analyses.

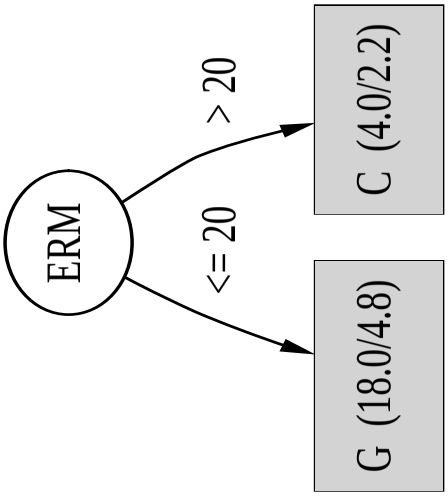
7.3 Discussion of results

A small decision tree that accurately describes the buying patterns of growers and consultants was created.

7.4 Conclusion

This data set had a class value that is already enumerated which makes it easy to use with machine learning. Both the results and the tree size were suitable for machine learning. However, the size of the data set may indicate bias in the results. A larger dataset may produce significantly different results. The error count may be contributed to an answer that was not consistent with other relationships formed with that particular attribute (GDD_calc).

Graph 1 : Result 1



8. PASTURE PRODUCTION

Data source Dave Barker
AgResearch Grasslands
Palmerston North

Report date December, 1995

Assigned to Stephen Garner, Research programmer

Data location
/home/ml/datasets/Summer1995/pasture_production/REPORT/results.arff

8.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To predict pasture production from a variety of biophysical factors. Vegetation and soil variables areas of grazed North Island hill country with different management (fertiliser application/stocking rate) histories (1973 - 1994) were measured and subdivided into 36 paddocks. Nineteen vegetation (including herbage production); soil chemical, physical and biological; and soil water variables were selected as potentially useful biophysical indicators.

Summary of original research results
Analysis of the results for these variables suggested that pasture botanical composition (especially content of high-fertility responsive grasses and of herbaceous weed species) and earthworm mass/unit area explained most of the variance of the produced data matrix. These variables were also highly correlated with herbage production, an indicator of likely economic performance.

8.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To predict pasture production from a variety of biophysical factors.

The original dataset has been sent but also includes a new variable, leaf phosphorous concentration (ug/g). The data provider requested that analysis be performed including this attribute as a separate test.

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
The dataset is small and contains no missing values. The factors present are only a representation of all the features required to describe pasture production.

Final dataset : 36 instances.

Attribute information
Number of attributes : 25 (24 actual attributes plus one predefined class).

The attributes are :

1. paddock_id (paddock identification number)
   type : integer
   values : 1.0 - 36.0
   missing : 0
2. pasture-prod-num (pasture production)
   type : integer
   values : 3067.0 - 11,904.0
   missing : 0
3. pasture-prod-class (class variable)
   type : enumerated
   values : LO, MED, HI
   missing : 0
4. Fertiliser
   type : enumerated
   values : LL, LN, HN, HH
   missing : 0
5. Slope
   type : integer
   values :7.0 - 33.0
   missing : 0
6. Aspect (deviation from north-west)
   type : integer
   values : 17.0 - 172.0
   missing : 0
7. OlsenP
   type : integer
   values : 5.0 - 42.0
   missing : 0
8. MinN
   type : integer
   values : 162.0 - 409.0
   missing : 0
9. TS
   type : integer
   values : 175.0 - 655.0
   missing : 0

10. Ca:Mg (Calcium Magnesium ratio)
    type : real
    values : 2.22 - 6.36
    missing : 0
11. LOM (g/100g)
    type : real
    values : 1.3 - 3.7
    missing: 0
12. NFIX mean
    type : real
    values : 0.012 - 0.229
    missing : 0
13. Eworms (main 3 spp - g/m^2)
    type : real
    values : 0.0 - 285.7
    missing : 0
14. Eworms (number of spp)
    type : integer
    values : 1.0 - 5.0
    missing : 0
15. KUnSat (mm/hr)
    type : real
    values : 4.4 - 57.3
    missing : 0
16. OM
    type : real
    values : 5.5 - 11.7
    missing : 0
17. Air-Perm
    type : real
    values : 2.63e-13 - 1.03e-11
    missing : 0
18. Porosity
    type : real
    values : 0.126 - 0.212
    missing : 0
19. HFRG% (mean)
    type : real
    values : 1.31 - 57.26
    missing : 0
20. legume-yield (kgDM/ha)
    type : real
    values : 26.4 - 642.7
    missing : 0
21. OSPP% (mean)
    type : real
    values : 1.19 - 19.07
    missing : 0

22. Jan-Mar-mean-TDR
    type : real
    values : 18.0 - 39.3
    missing : 0
23. Annual-Mean-Runoff (mm)
    type : real
    values : 578.9 - 893.5
    missing : 0
24. root-surface-area (m^2/m^3)
    type : real
    values : 56.2 - 520.6
    missing : 0
25. Leaf-P (ppm)
    type : real
    values : 1162.0 - 3993.0
    missing : 0

Class information

The datasets contained one class (pasture-prod-class) which has three class values :

1. LO (12 lowest values)
2. MED (12 intermediate values)
3. HI (12 highest values)

The class values were assigned arbitrarily by the data provider.

The distribution of the classes are :

1. LO 12
2. MED 12
3. HI 12

Analysis with 1Rw produced the following results :
default class : 'LO' with 12/36
best hypothesis : 'pasture-prod-num' with 100% accuracy

Data analysis procedure	**Results**
Two original attributes were altered; 'Ca:Mg' was multiplied by 20 and numbers in 'AirPerm' were converted from scientific notation to decimal notation.	

1. C4.5 was run using 'pasture-prod-class' as the class with all attributes except : 'paddock_id', 'pasture-prod-num' 'Air-Perm' 'Leaf-P'	1. The decision and simplified trees were : *size : 9 with 2.8% errors.* dominant attributes : 'HFRG%mean', 'Leaf-P', 'legume-yield' 'Eworms-main-3'
2. C4.5 was run using 'pasture-prod-class' as the class with all attributes except : 'paddock_id', 'pasture-prod-num' 'Air-Perm' ('Leaf-P' was included this time)	2. The decision and simplified trees were : *size : 9 with 2.8% errors.* dominant attributes : 'HFRG%mean' 'NFIX-mean' 'Eworms-main-3' 'legume-yield'.

The results are presented at the end of this analyses.

8.3 Discussion of results

Small decision trees were produced that describe possible ways of classifying pasture production. It is interesting that when 'Leaf-P' is removed, the resulting tree is quite similar with 'NFIX-mean' almost replacing 'Leaf-P' with 'Eworms-main-3' and 'legume-yield' swapping positions.

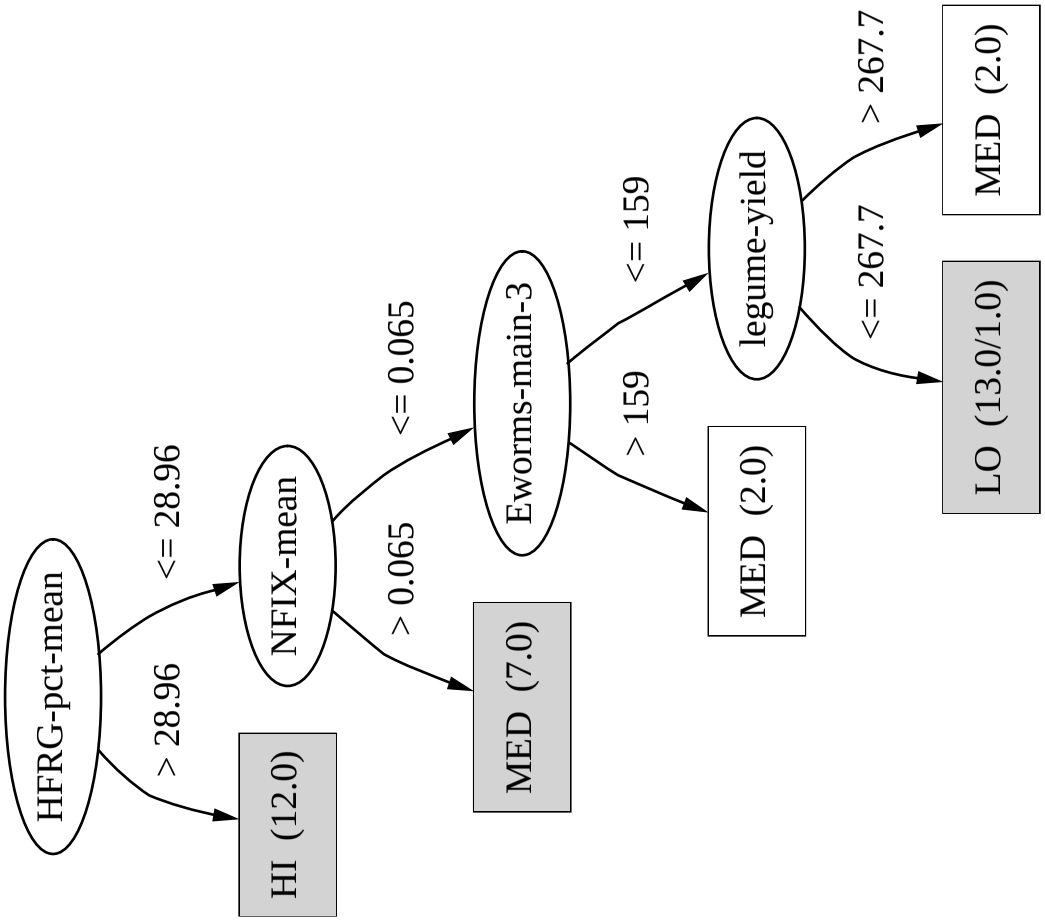
8.4 Conclusion

The type of data was suitable for machine learning analyses, but the size of the dataset was too small to get any results that can be definitely accurate. The results that have been found could be too pronounced. Analysis with more instances could create a different result.

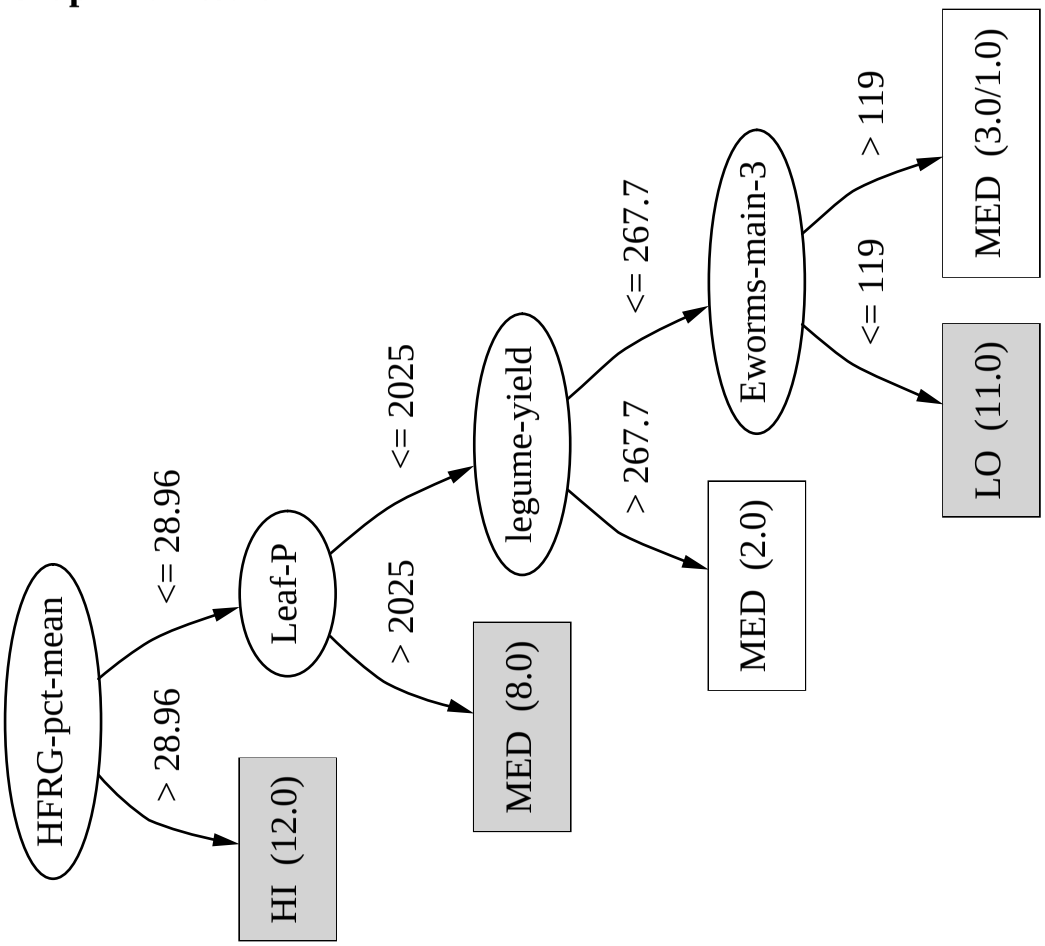
Follow up notes
Dave Barker felt that the classification was good.

Reference : Lambert M. G., Barker D. J., MacKey A. D., Springett J. A. (1995) Biophysical indicators of sustainability of North Island hill pasture systems.*Proceeding of the New Zealand Grassland Association*, Vol 57.

Graph 1 : Result 1



Graph 2 : Result 2



9. PEA SEED COLOUR

Data source John McCallum
Crop & Food Research
Christchurch

Report date January 1996

Assigned to Stuart Yeates, Research programmer

Data location /home/ml/datasets/Summer1995/pea/REPORT/original.arff
/home/ml/datasets/Summer1995/pea/REPORT/results.arff

9.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To optimise a pea transformation method (ie. putting foreign DNA into them) by working out the optimum stage for taking pods. An objective method for getting the best pods is sought. This is based on going out to the field and getting a big pile of pods and classifying them into either 'yes' or 'no' piles. Video image analyses is not suitable and the pods should not have to be opened.

Summary of original research results
A dataset has been produced from the methods used above. Any analysis results are not known at this particular time.

9.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To classify pea pods as 'good' or 'bad' using attributes about the visual aspects of the pods. The results from machine learning will be compared to the more conventional analysis procedures such as statistical analysis. The results gained from machine learning will be used to extract information anout the rules used by judges to classify things (like peapods).

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information

There are no missing values. The factors present are only a representation of all the features required to describe classification or pea seed colour.

Original dataset : 51 instances.
Final dataset : 51 instances.

Attribute information
Number of attributes : 15 (14 actual attributes plus one predefined class).

The attributes are :

1. Code (id number for each pod)
   type : enumerated
   values : 0 - 50
   missing : 0
2. Status (visual classification)
   type : enumerated
   values : 1 (yes), 0 (too young)
   missing : 0
3. Shape
   type : real
   values : 67.86 - 81.16
   missing : 0
4. Height
   type : real
   values : 163.0 - 335.0
   missing : 0
5. Length
   type : real
   values : 278.0 - 553.61
   missing : 0
6. Breadth
   type : real
   values : 82.16 - 167.26
   missing : 0
7. Width (pod width)
   type : real
   values : 181.0 - 471.0
   missing : 0
8. Diam45 (pod diameter)
   type : real
   values : 210.01 - 542.35
   missing : 0
9. Diam135 (pod diameter)
   type : real
   values : 70.71 - 388.91
   missing : 0
10. Red (tristimulus computer RGB measures of pod colour)
    type : real
    values : 166.64 - 178.81
    missing : 0

11. Green (tristimulus computer RGB measures of pod colour)
    type : real
    values : 46.75 - 59.52
    missing : 0
12. Blue (tristimulus computer RGB measures of pod colour)
    type : real
    values : 46.75 - 59.52
    missing : 0
13. Y (transform of red)
    type : real
    values : 133.831 - 149.339
    missing : 0
14. U (transform of green)
    type : real
    values : -99.7988 - -81.3821
    missing : 0
15. V (transform of blue)
    type : real
    values : 21.1393 - 29.4712
    missing : 0

Class information

'*Status'* is only one class value :

1. 0 (too young)
2. 1 (yes)

The distribution of the classes are :

1. 16
2. 35

Analysis with 1Rw indicated the following results :

default class : '1' with 35/51
best hypothesis : 'Shape' with 80% accuracy.

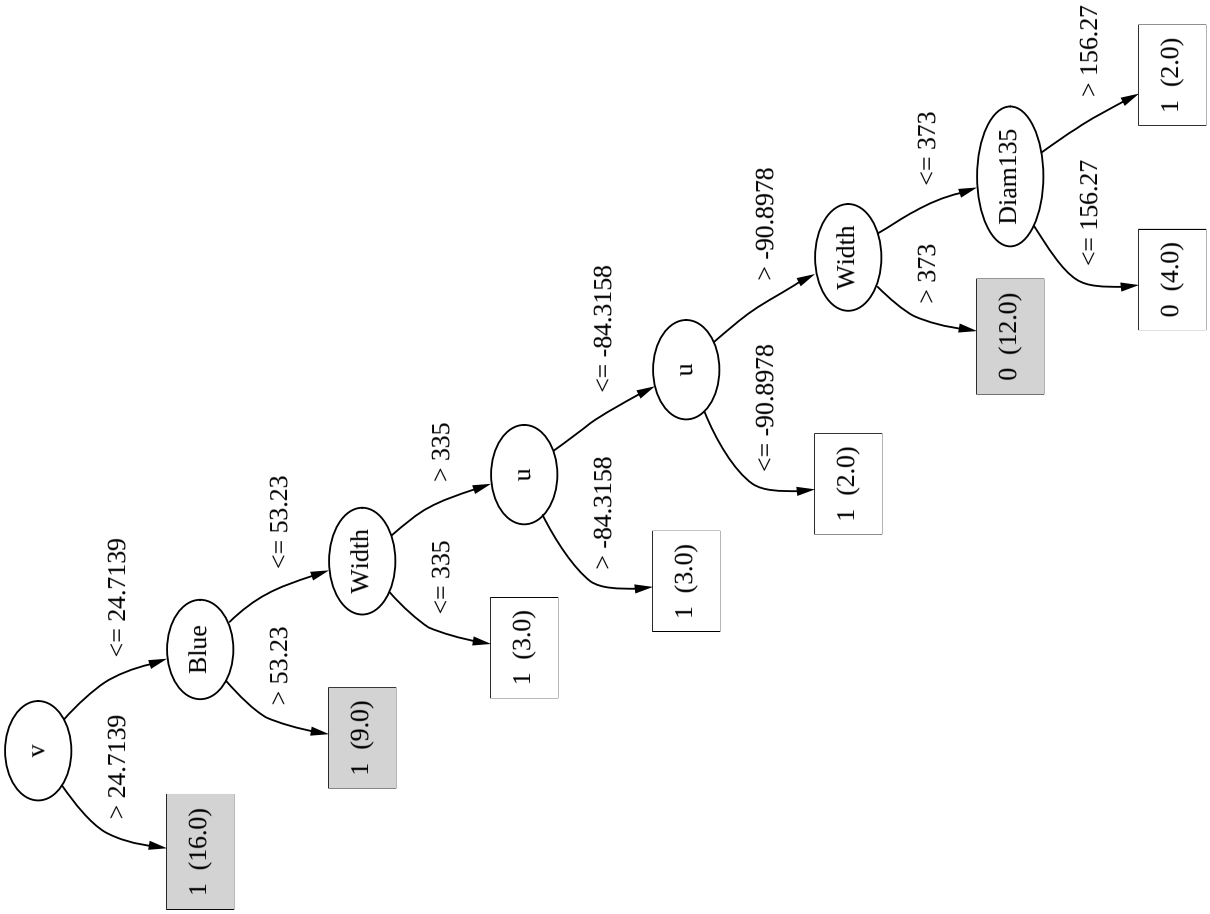
Data analysis procedure	**Results**
1. C4.5 was run using 'Status' as the class and all attributes except 'Code'.	1. The simplified decision tree was : *size : 7 nodes with 18% errors* dominant attributes are : 'v' 'Blue' 'Width' 'u' 'Diam'

The results are presented at the end of this analyses.

9.3 Conclusion

C4.5 was able to find a small decision tree with few errors from this data, mainly because it has a clearly defined goal that is approachable from a machine learning point of view. Also, the class attribute was given as an enumerated (binary) value so (no false classes needed to be created). There was no missing data, which made the level of noise and errors low.

Graph 1 : Result 1



10. SLUGS

Data source Neil Cox
AgResearch Ruakura
Hamilton

Date received October 1995

Assigned to Colleen Burrows, Research programmer
James Littin, Research programmer

Data location /home/ml/datasets/Summer1995/slug/REPROT/origianl.arff
/home/ml/datasets/Summer1995/slug/REPORT/s9_results.arff
/home/ml/datasets/Summer1995/slug/REPORT/s10_results.arff

10.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To detect in the data when a new bucket of slugs was used and when the weight measurements were taken wrong.

Summary of original research results
The results of the original research are currently unknown.

10.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To detect in the data when a new bucket of slugs was used and when the weight measurements were taken wrong. There were two problems in the dataset that we tried to detect with the machine learning tools :

1. Two buckets of slugs were used; larger slugs were supposedly measured first. Could we detect the trend of slugs getting smaller gradually and then an abrupt increase when a new bucket was used
2. We were informed that for part of the experiment there was supposedly some muck on the scale to make the values read higher than the actual slug weight

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
The data provided by the dataset contained no errors or obvious noise.

Original dataset : 100 instances.
Final datasets : 100 instances.

Attribute information
Number of attributes : two.

The attributes are :

1. Length
   type : real
   values : 10 - 94
   missing : 0
2. Weight
   type : real
   values : 0.02-13.62
   missing : 0

A number of derived attributes were created. These attributes are :

3. No (order of slug taken)
   type : integer
   values : 0.0 - 99.0
   missing : 0
4. Weight_enum (division at mean)
   type : enumerated
   values : heavy, light
   missing : 0
5. Weight_enum2 (division at 0.6 and 1.4of normalised weight)
   type : enumerated
   values : verylight, light, heavy
   missing : 0
6. Length_enum (division at 0.6 and 1.4 of normalised length)
   type : enumerated
   values : short, medium, long
   missing : 0
7. w/13 (weight/length3, density)
   type : real
   values :
   missing :
8. w/13_enum (density)
   type : enumerated
   values : small, big
   missing : 0
9. w/13_enum2 (density)
   type : enumerated
   values : small, medium, large
   missing : 0

Class information

A number of class values were derived from the given dataset.

weight_enum

1. heavy
2. light

weight_enum2

1. heavy
2. light
3. verylight

length_enum

1. long
2. medium
3. short

w/13_enum

1. small
2. big

w/13_enum2

1. small
2. medium
3. large

The distribution of the classes are :

weight_enum

1.	heavy	31
2.	light	69

weight_enum2

1.	heavy	28
2.	light	8
3.	verylight	64

length_enum

1.	long	26
2.	medium	40
3.	short	34

w/13_enum

1.	small	66
2.	big	34

w/13_enum2

1.	small	18
2.	medium	68
3.	large	14

Analysis with 1Rw produced the following results :

weight_enum
default class : 'light' with 69/100
best hypothesis : 'weight' with 100% accuracy

weight_enum2
default class : 'verylight' with 64/100
best hypothesis : 'weight' with 100% accuracy

length_enum
default class : 'medium' with 40/100
best hypothesis : 'length' with 100% accuracy

w/13_enum
default class : 'small' with 66/100
best hypothesis : 'length' with 76% accuracy

w/13_enum2
default class : 'medium' with 68/100
best hypothesis : 'length' with 75% accuracy.

Data analysis procedure	**Results**
The attribute 'weight' was normalised to make the mean at 1.0 and a new attribute, 'weight_enum' was created with a division at the mean.	
1. C4.5 was used with 'weight_enum' as the class and the 'No' (the order in which the slugs were measured) as the only attribute.	1. The simplified decision tree was *size 13 (nodes) with 7% errors* The decisions follow the trends found in a graph of weight vs slug number.
A new attribute, 'weight_enum2', was created which divided the normalised weight attribute at 0.6 and 1.4.	
2. C4.5 was run with this as the class and 'No' as the attribute.	2. The simplified decision tree was size : 9 with 15% errors dominant attributes were the lowest and highest classes.

	This followed the curve, not quite as closely as the previous tree, but it was still able to predict the main point of where the new bucket was used (at about slug #67).
Length was then used to try to detect when the bucket was changed. A normalised attribute for length was created which was discretised at the mean. 3. C4.5 was run using this attribute, 'length_enum', as the class and 'No' as the attribute.	3. A simplified decision tree was size : 21 with 15% errors This followed the curves of the graph of length vs slug number, but it is a much larger tree than those found when weight was used. Again, it found the instance where the bucket was changed was at slug 67.
The problem of detecting when the muck arrived on the scales was approached by creating a new attribute of density, 'w/l3' which was discretised at the mean of 0.61. 4. C4.5 was used with 'w/l3_enum' as the class and 'No' as the only attribute. A second density attribute was created that split the attribute into three classes (small (less than 0.4), medium (less than 0.9), large (greater than 0.9).	4. The simplified decision tree was : size : 25 with 13% errors
5. C4.5 was run with this attribute as the class and 'No' as the only other attribute.	5. The simplified decision tree was size : 11 with 22% errors. It shows that the density starts out small, and increases to medium and large for the rest of the values which could be a sign of detection of muck on the scales.

The results are presented at the end of this analyses.

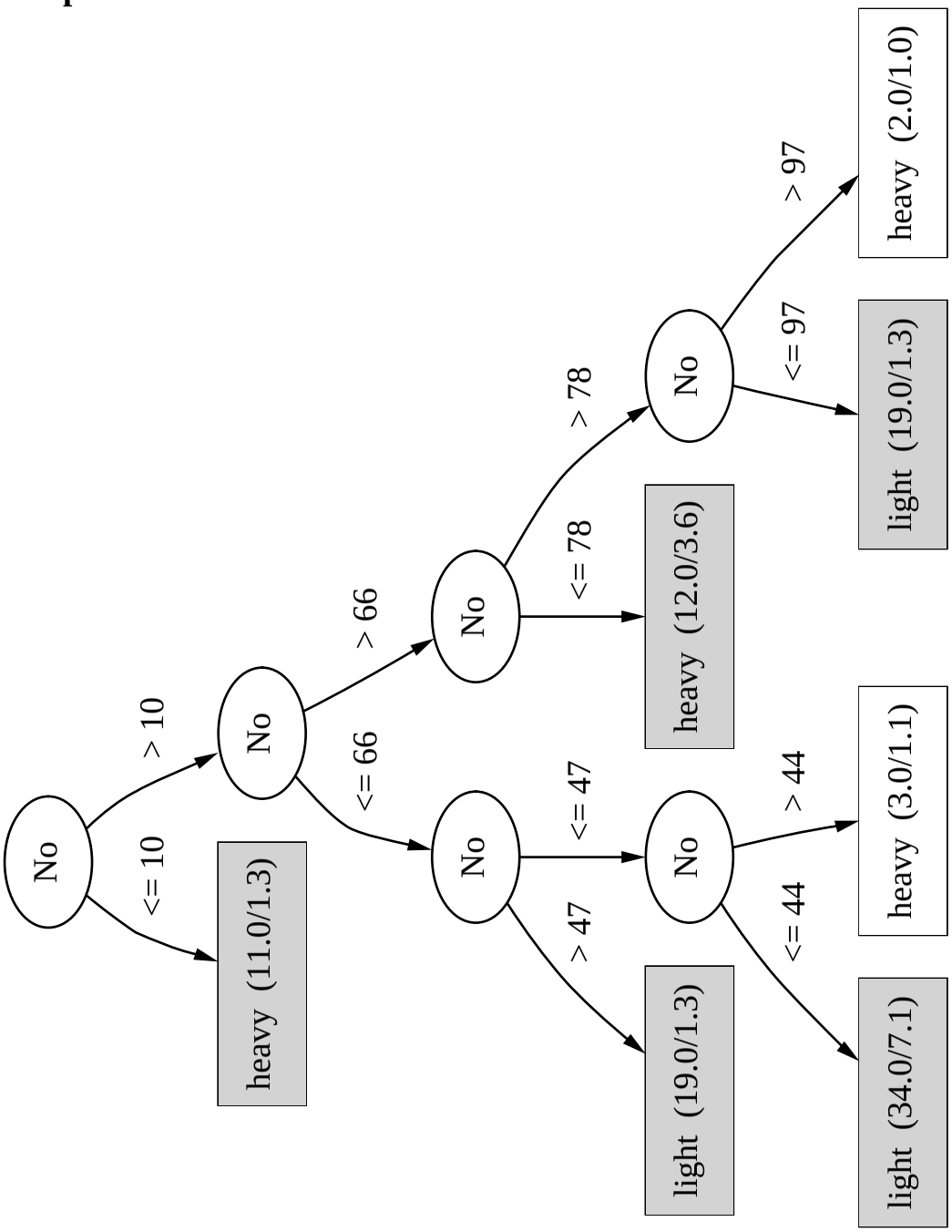
10.3 Discussion of results

The decision trees were larger than expected, but they did show what is present in the graphs ie. C4.5 was able to pick out where the changes in data occurred.

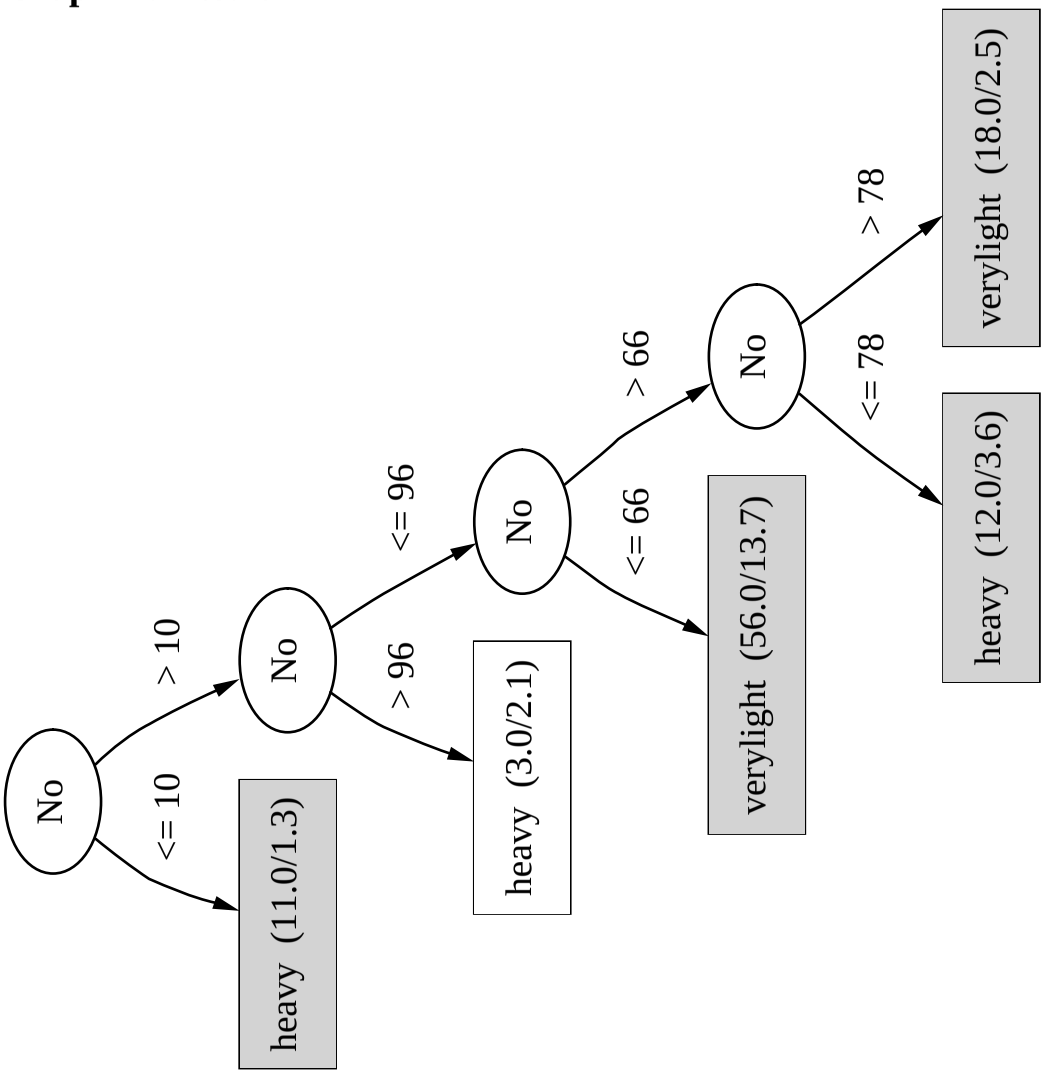
10.4 Conclusion

Although we were able to detect changes in the data, it is probably much easier to plot the values on a graph and use that for analysis. Since there are only two attributes, graphical or statistical analysis would not be difficult.

Graph 1 : Result 1



Graph 2 : Result 2



11. SQUASH HARVEST

Data source Winna Harvey
Crop & Food Research
Christchurch

Report date February, 1996

Assigned to Kirsten Thomson, Research programmer

Data location /home/ml/datasets/Summer1995/Squash/REPORT/stored_results.arff
/home/ml/datasets/Summer1995/Squash/REPORT/unstored_results.arff

11.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives

The purpose of the experiment was to determine the changes taking place in squash fruit during maturation and ripening so as to pinpoint the best time to give the best quality at the market place (Japan). The squash is transported to Japan by refrigerated cargo vessels (14^{o}C) and takes three to four weeks after harvesting to reach the market. Evaluations were carried out at a stage representing the quality inspection stage prior to export and also at the stage it would reach on arriving at the market place.

Summary of original research results

These are unknown at the present time.

11.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective

To determine which pre-harvest variables contribute to good tasting squash after different periods of storage time. This is determined by whether a particular squash fell into the categories of unacceptable, acceptable, or excellent - based on the measure of acceptability.

Dataset description
The dataset descriptions include instance, attribute, and class information. The original dataset contained four different categories for which data was collected. It was more meaningful to divide the original dataset into these four datasets.

The four datasets are :
1. stored training - stored fruit with an acceptability (sensory) rating
2. unstored training - unstored fruit with an acceptability (sensory) rating
3. stored testing - stored fruit without an acceptability (sensory) rating
4. unstored testing - unstored fruit without an acceptability (sensory) rating.

Only the training sets have been analysed further. Inadequate information is provided to perform analysis on the testing sets.

Instance information
A high number of instances contained missing values and noise. The factors present are only a representation of all the features required to describe

Original dataset : 261 instances.
Unstored training : 52 instances.
Stored training : 52 instances.

Attribute information
Number of attributes : (original dataset) 24 attributes.
Number of attributes : (unstored training dataset) 24 attributes.
Number of attributes : (stored training dataset) 24 attributes.

The attributes are :

Stored training

1. site
   type : enumerated
   values : P, HB, LINC
   missing : 0
2. daf (days after flowering)
   type : enumerated
   values : 30, 40, 50, 60, 70
   missing : 0
3. fruitno (individual number of the fruit)
   type : enumerated
   values : 1 - 23
   missing : 0
4. weight (of whole fruit in grams)
   type : real
   values : 1156.0 - 2872.0
   missing : 0
5. storewt (weight of fruit after storage)
   type : real
   values : 1067.0 - 2607.0
   missing : 0

6. pene (penetrometer - indication of maturity at harvest)
   type : integer
   values : 2.0 - 11.0
   missing : 0
7. solids_% (a dry matter test)
   type : integer
   values : 13.0 - 30.0
   missing : 0
8. brix (a refractometer measurement - an indication of sweetness and/or ripeness)
   type : integer
   values : 7.0 - 15.0
   missing : 0
9. a* (the a* co-ordinate of the Hunterlab L* a* b* notation of colour measurement)
   type : integer
   values : 10.0 - 29.0
   missing : 0
10. egdd (the heat accumulation above a base of 8ºC from emergence of the plant to harvest of the fruit)
    type : real
    values : 601.0 - 953.0
    missing : 0
11. fgdd (the heat accumulation above a base of 8ºC from flowering to harvest)
    type : real
    values : 190.0 - 542.0
    missing : 0
12. groundspot_a* (number indicating colour of skin where the fruit rested on the ground)
    type : integer
    values : 7.0 - 19.0
    missing : 1
13. glucose (mg/100 grams fresh weight)
    type : integer
    values : 4.0 - 24.0
    missing : 1
14. fructose (mg/100 grams fresh weight)
    type : integer
    values : 4.0 - 20.0
    missing : 1
15. suctose (mg/100 grams fresh weight)
    type : integer
    values : 4.0 - 46.0
    missing : 1
16. total (mg/100 grams fresh weight)
    type : integer
    values : 29.0 - 70.0
    missing : 1

17. glucose+fructose (mg/100 grams fresh weight)
    type : integer
    values : 9.0 - 44.0
    missing : 1
18. starch (mg/100 grams fresh weight)
    type : integer
    values : 27.0 - 175.0
    missing : 1
19. sweetness (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 319.0 - 955.0
    missing : 0
20. flavour (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 415.0 - 1008.0
    missing : 0
21. dry/moist (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 188.0 - 949.0
    missing : 0
22. fibre (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 68.0 - 686.0
    missing : 0
23. acceptability (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 557.0 - 1224.0
    missing : 0
24. heat_input_emerg (heat emergence after harvest)
    type : real
    values : 721.0 - 1087.0
    missing : 0
25. heat_input_flower (heat input before flowering)
    type : real
    values : 386.0 - 738.0
    missing : 0

Unstored training

1. site
    type : enumerated
    values : P, HB, LINC
    missing : 0
2. daf (days after flowering)
    type : enumerated
    values : 30, 40, 50, 60, 70
    missing : 0

3. fruitno (individual number of the fruit)
   type : enumerated
   values : 1 - 27
   missing : 0
4. weight (of whole fruit in grams)
   type : real
   values : 1064.0 - 2523.0
   missing : 0
5. storewt (weight of fruit after storage)
   type : real
   values : N/A
   missing : 0
6. pene (penetrometer - indication of maturity at harvest)
   type : integer
   values : 3.0 - 11.0
   missing : 0
7. solids_% (a dry matter test)
   type : integer
   values : 13.0 - 24.0
   missing : 0
8. brix (a refractometer measurement - an indication of sweetness and/or ripeness)
   type : integer
   values : 5.0 - 14.0
   missing : 0
9. a* (the a* co-ordinate of the Hunterlab L* a* b* notation of colour measurement)
   type : integer
   values : 1.0 - 24.0
   missing : 0
10. egdd (the heat accumulation above a base of 8ºC from emergence of the plant to harvest of the fruit)
    type : real
    values : 601.0 - 953.0
    missing : 0
11. fgdd (the heat accumulation above a base of 8ºC from flowering to harvest)
    type : real
    values : 190.0 - 542.0
    missing : 0
12. groundspot_a* (number indicating colour of skin where the fruit rested on the ground)
    type : integer
    values : -9.0 - 17.0
    missing : 2
13. glucose (mg/100 grams fresh weight)
    type : integer
    values : 1.0 - 17.0
    missing : 6

14. fructose (mg/100 grams fresh weight)
    type : integer
    values : 1.0 - 15.0
    missing : 6
15. suctose (mg/100 grams fresh weight)
    type : integer
    values : 3.0 - 39.0
    missing : 6
16. total (mg/100 grams fresh weight)
    type : integer
    values : 15.0 - 60.0
    missing : 6
17. glucose+fructose (mg/100 grams fresh weight)
    type : integer
    values : 3.0 - 32.0
    missing : 6
18. starch (mg/100 grams fresh weight)
    type : integer
    values : 44.0 - 234.0
    missing : 6
19. sweetness (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 225.0 - 957.0
    missing : 0
20. flavour (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 350.0 - 1002.0
    missing : 0
21. dry/moist (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 151.0 - 882.0
    missing : 0
22. fibre (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 119.0 - 416.0
    missing : 1
23. acceptability (taste panel score - mean of 8 panellists; score out to 1500)
    type : integer
    values : 345.0 - 1224.0
    missing : 0
24. heat_input_emerg (heat emergence after harvest)
    type : real
    values : 671.0 - 1023.0
    missing : 0
25. heat_input_flower (heat input before flowering)
    type : real
    values : 260.0 - 612.0
    missing : 0

Class information

Only one class was created, based on acceptability.

1. not_acceptable less than 750
2. acceptable more than 750 and less than and including 1000
3. excellent more than 1000 and less than and incuding 1500

The distribution of the classes for both training files are (these are the files that we worked with) :

Stored training

1.	not_acceptable	8
2.	acceptable	21
3.	excellent	23

Unstored training

1.	not_acceptable	24
2.	acceptable	24
3.	excellent	4

Analysis with 1Rw has provided the following results :

Stored training

default class : 'excellent' with 23/52

best hypothesis : 'total' with 69% accuracy

Unstored training

default class : 'acceptable' with 24/52

best hypothesis : 'sweetness' with 79% accuracy.

Data analysis procedure	**Results**
1. C4.5 was run with new_acceptability' as the class and with all other attributes in the stored training set	1. The simplified decision tree was : size : 18 (nodes) with 5.8% errors dominant attributes : pene solids_%
2. C4.5 was run with 'new_acceptability' as the class and with all other attributes in the unstored training set	2. The simplified decision tree was : size : 11 (nodes) with 3.8% errors dominant attributes : pene solids_%

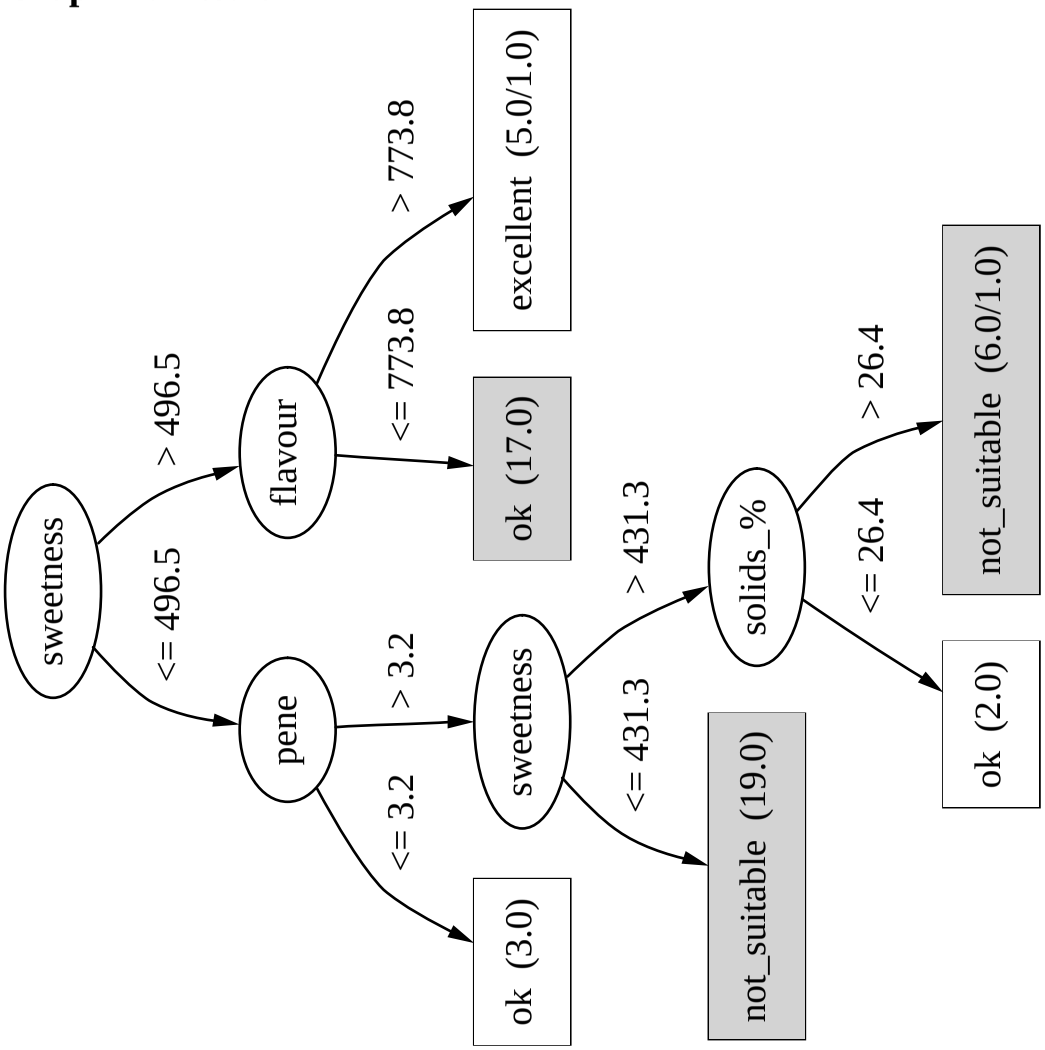
3. C4.5 was run with 'new_acceptability' as the class but excluding the sensory attributes for the stored training set	3. The simplified decision tree was : size : 18 (nodes) with 5.8% errors dominant attributes : sweetness flavour pene
4. 1Rw was run with 'new_acceptability' as the class but excluding the sensory attributes for the unstored training set	4. The simplified decision tree was : size : 13 (nodes) with 9.6% errors dominant attributes : a* fructose, solids_%

The results are presented at the end of this analyses.

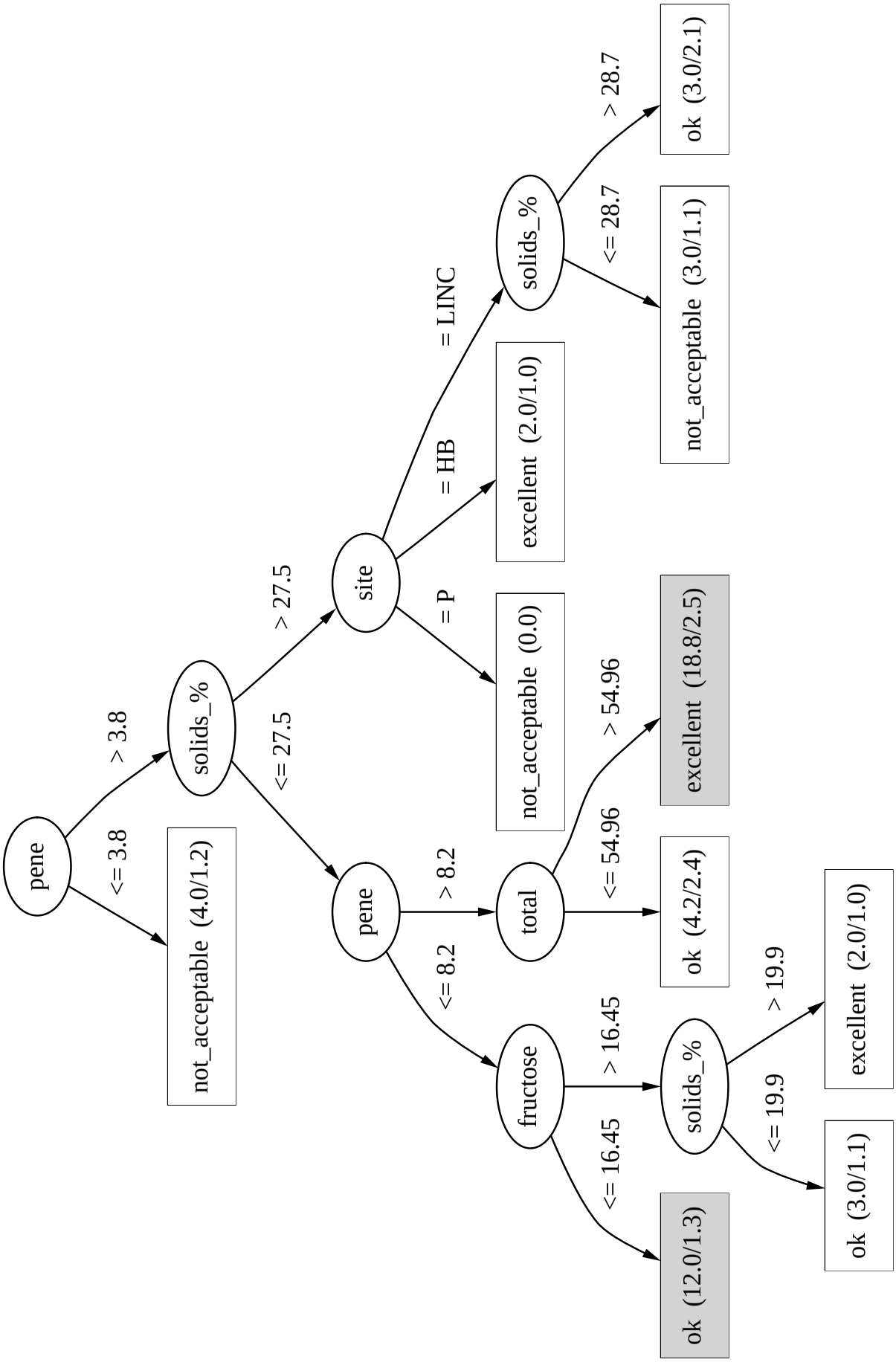
11.3 Conclusion

Both the size of the trees and the number of errors are small. The structure of the trees are logical and easy to follow. The type of data and the results gained from these indicate that this dataset was a good one for machine learning. However, the results were not what were required by the researcher. In this case the ability to gain the required answers was not obtainable. The data may need to be formatted and collected in a way that is more suitable to machine learning.

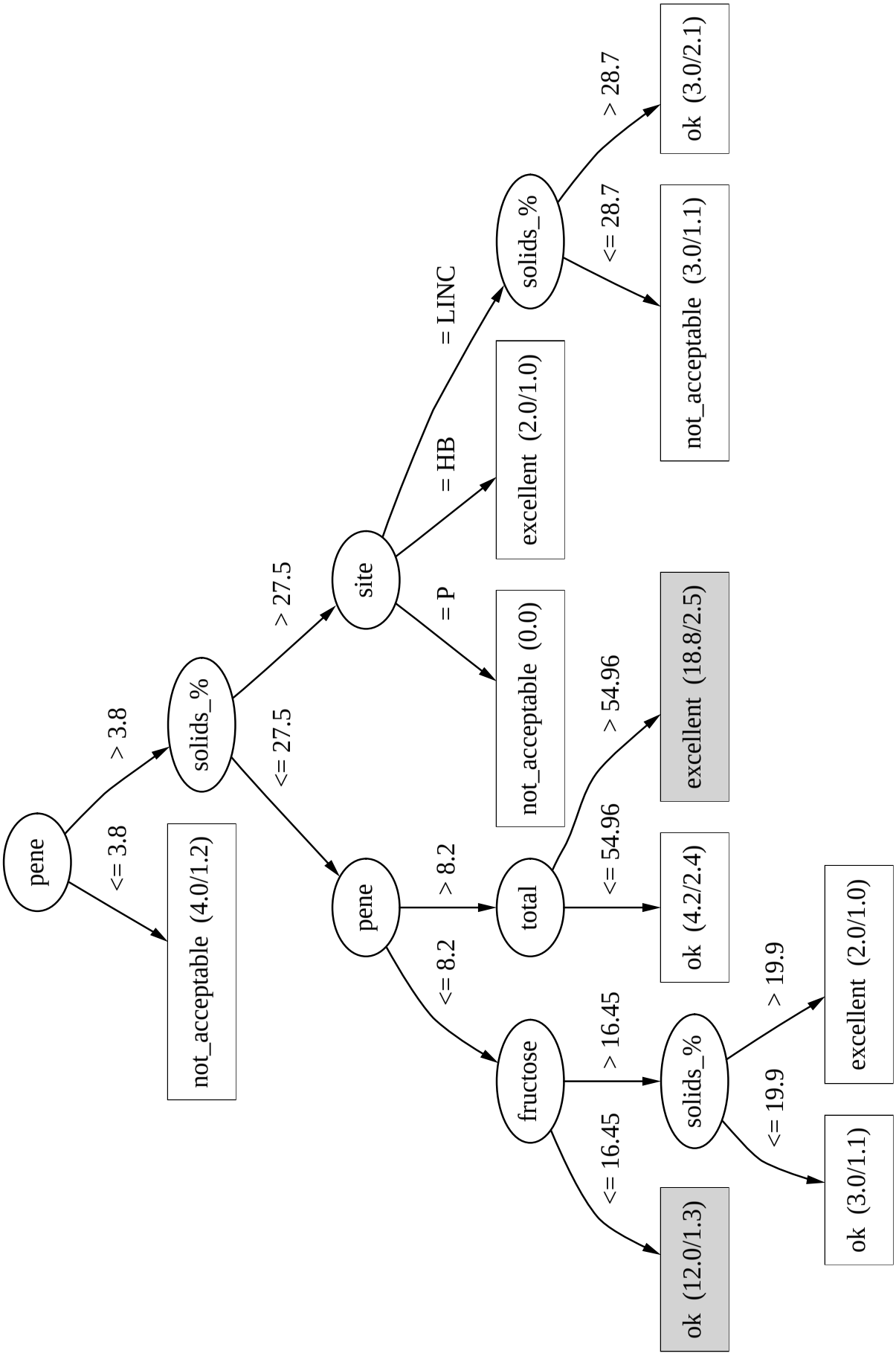
Graph 1 : Result 2



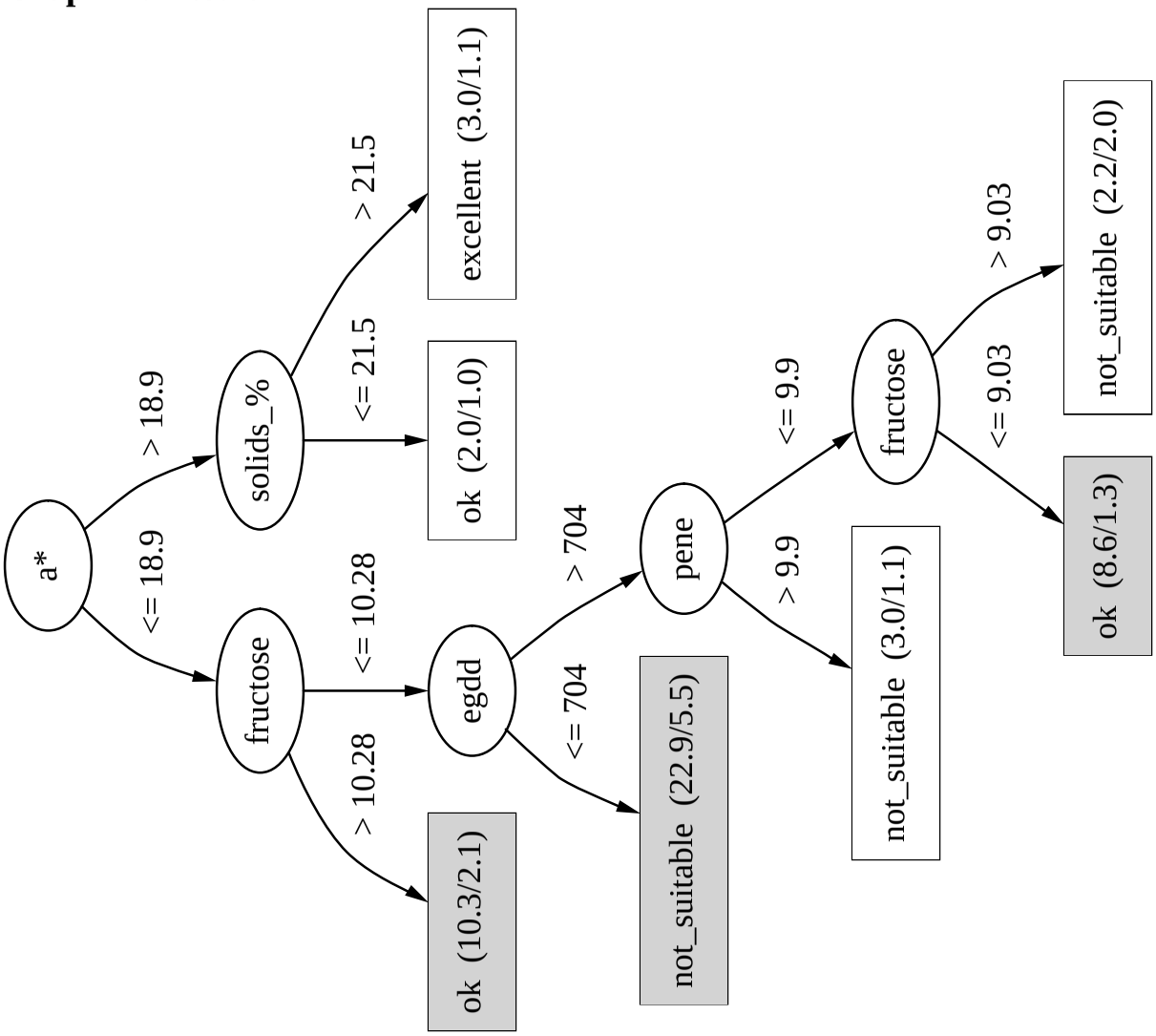
Graph 2 : Result 3



Graph 3 : Result 1



Graph 4 : Result 4



12. VALLEY CATEGORIES

Data source Liddy Bakker
Marine and Geoscience
University of Waikato
Hamilton

Report date November 1995

Assigned to Stephen Garner, Research programmer

Data location /home/ml/datasets/Summer1995/valley/REPORT/original.arff
/home/ml/datasets/Summer1995/valley/REPORT/results.arff

12.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To find and describe clusters of similar sections of valley.

Summary of original research results
The original results are currently unknown.

12.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To find and describe clusters of similar sections of valley.

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
A high number of instances contained missing values and noise. The factors present are only a representation of all the features required to describe valley categories.

Original dataset : 878 instances.
Final dataset : 878 instances.

Attribute information
Number of attributes : 15 (14 actual attributes plus one predefined class).

The attributes are :

1. BlockNumber
   type : integer
   values : 2.0 - 1,539.0
   missing : 0
2. PercentValley
   type : real
   values : 0.0 - 100.0
   missing : 0
3. FS1+FS7
   type : real
   values : 0.0 - 100.0
   missing : 0
4. FS2+FS8
   type : real
   values : 0.0 - 100.0
   missing : 0
5. FS3+FS9
   type : real
   values : 0.0 - 100.0
   missing : 0
6. GS1+SS1+VS1+GS7+SS7+VS7
   type : real
   values : 0.0 - 100.0
   missing : 0
7. GS2+SS2+VS2+GS8+SS8+VS8
   type : real
   values : 0.0 - 100.0
   missing : 0
8. GS3+SS3+VS3+GS9+SS9+VS9
   type : real
   values : 0.0 - 100.0
   missing : 0
9. GS4+SS4+VS4
   type : real
   values : 0.0 - 100.0
   missing : 0
10. GS5+SS5+VS5
    type : real
    values : 0.0 - 100.0
    missing : 0
11. GS6+SS6+VS6
    type : real
    values : 0.0 - 100.0
    missing : 0
12. FS4
    type : real
    values : 0.0 - 100.0
    missing : 0

13. FS5
    type : real
    values : 0.0 - 100.0
    missing : 0
14. FS6
    type : real
    values : 0.0 - 100.0
    missing : 0
15. class
    type : enumerated
    values : 0, 1, 2, 3, 4, 5, 6
    missing : 0

Class information

For the dataset there are six class values.

1. 0
2. 1
3. 2
4. 3
5. 4
6. 5

The distribution of the classes are :

1. 0 213
2. 1 197
3. 2 125
4. 3 126
5. 4 108
6. 5 38

Analysis with 1Rw produced the following results :

default class : '0' with 213/878
best hypothesis : 'FS2+FS8' with 58% accuracy.

Data analysis procedure	**Results**
1. Autoclass was used to cluster the blocks into distinct classes.	1. It came up with seven classes. Ray Littler (Applied Statistician at the University of Waikato) also came up with a similar number of classes (six classes) but one of these classes contained almost all of the instances and the other classes contained only a few instances.

2. C4.5 was run using 'class' as class and with all attributes	2. The rules determine from the tree had a *66/34 split on the data.* The test set had about 65% accuracy of the rules on the test set.
Further data on relative positioning of blocks was obtained and used to plot the blocks and which class they belong in. There are definite areas throughout the map where groups of one class dominate. Some of the 36 classes were combined to produce only 12 classes.	
3. This derived data was run on Autoclass	3. Seven new categories of data were produced. These were then used to make a new rule set and plotted on the map. The map produced different areas grouped together by similarities in valley category. This indicates that differences exist between the categories produced when 36 classes were used and when 12 classes were used.

The results are presented at the end of this analyses.

12.3 Discussion of results

Because of the nature and objectives of the experiment, the results from the data were inconclusive.

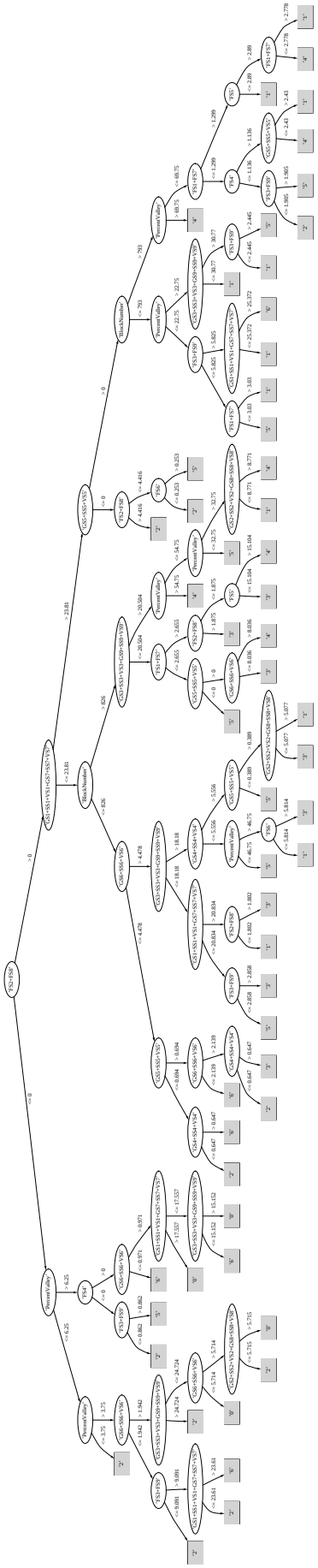
12.4 Conclusion

This dataset may have been suitable for machine learning. However, the lack of a defined goal meant that we were unable to come up with feasible solutions. Also, the type of data required strong domain knowledge. This could only come from the researcher. Analysis by machine learning often require a well-defined question to answer and strong interaction with the researcher to determine which results are suitable, and which can be built on.

Follow up notes

This is an on-going project, where the researcher will be required to have more input.

Graph 1 : Result 2



13. VENISON BRUISING

Data source Jenny Jago
Animal Behaviour and Welfare
AgResearch
Ruakura Research Centre, Hamilton

Report date December 1995

Assigned to Stuart Yeates, Research programmer

Data location /home/ml/datasets/Summer1995/venison/REPORT/original.arff
/home/ml/datasets/Summer1995/venison/REPORT/results.arff.gz

13.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To find the factors contributing most to bruising of deer carcasses when transported to to be slaughtered.

Summary of original research results
These results are currently unknown.

13.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To find the factors contributing most to bruising of deer carcass. Examine the data to determine if the distance travelled contributes to bruising.

Dataset description
The dataset descriptions include instance, attribute, and class information.
The distance attribute has a lot of missing values. The dataset is large with many attributes.

Instance information
A high number of instances contained missing values and noise. The factors present are only a representation of all the features required to describe deer bruising.

Original dataset : 21,454 instances.
Final dataset : 21,448 instances.

Attribute information
Number of attributes : 23 (21 actual attributes plus 2 predefined classes).

The attributes are :

1. docket (set of slaughter)
   type : integer
   values : 1244.0 - 3233.0
   missing : 0
2. month (the month of acquistion of the animal by the slaughter house)
   type : enumerated
   values : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
   missing : 0
3. day (the month of acquistion of the animal by the slaughter house)
   type : enumerated
   values : 1 - 31
   missing : 0
4. farm (the farm the deer came from)
   type : integer
   values : 1.0 - 501.0
   missing : 0
5. distance travelled (km)
   type : integer
   values : 24.0 - 434.0
   missing : 692
6. carrier
   type : enumerated
   values : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
   missing : 0
7. mixsex (single sex or mixed sex)
   type : enumerated
   values : 1, 2, 3, 4
   missing : 0
8. mixlot
   type : enumerated
   values : 1, 2, 3
   missing : 0
9. bookid (animal id)
   type : integer
   values : 1.0 - 21,500.0
   missing : 0
10. slaughter-order (order of slaughter for each docket)
    type : integer
    values : 1.0 - 99.0
    missing : 8
11. sex
    type : enumerated
    values : 1, 2
    missing : 5

12. lot-order
    type : integer
    values : 1.0 - 155.0
    missing : 6
13. hcw (hot carcass weight)
    type : real
    values : 0.7 - 663.0
    missing : 55
14. grade
    type : enumerated
    values : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
    missing : 7
15. bruising
    type : enumerated
    values : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
    missing : 5
16. other-damage
    type : enumerated
    values : 0, 1, 2, 3, 4, 5, 6, 7
    missing : 6
17. gr (measure of fatness)
    type : enumerated
    values : 0 - 43
    missing : 215
18. sltr-day (slaughter day)
    type : enumerated
    values : 1 - 31
    missing : 5
19. sltr-month (slaughter month)
    type : enumerated
    values : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
    missing : 5
20. other_damage_code
    type : enumerated
    values : none, deformed,..., other
    missing : 0
21. day-of-year
    type : integer
    values : 3.0 - 355.0
    missing : 0
22. sltr-day-of-year
    type : integer
    values : 3.0 - 355.0
    missing : 5
23. days_held
    type : integer
    values : -13.0 - 3.0
    missing : 5

Class information

A number of attributes were used as the class. Bruising(1), Bruising(2), calculated from the bruising attribute, and Grade were all used as classes.

Bruised

1. none
2. bruised

Bruise_Code

1. none
2. DH1
3. DM
4. DF1
5. Old_Scar_Break
6. Severe_Bruise
7. Condemned_Severe_Bruise
8. Emaciated
9. More-Than-One-Bruise
10. DH2
11. DF2
12. Broken-Bone-Haemorage
13. Condemned-DOA
14. Condemned-Emaciated
15. Condemned-Dead-Dying

Grade_Code :

1. AP5
2. AP3
3. AP4
4. other
5. AD
6. AF1
7. AP2
8. AP1
9. unknown
10. TBR
11. AM
12. AF2

The distribution of the classes are :

Bruised

1.	none	19,911
2.	bruised	1,543

Bruise_Code

1.	none	19,911
2.	DH1	1,036
3.	DM	96
4.	DF1	85
5.	Old_Scar_Break	79
6.	Severe_Bruise	15

7.	Condemned_Severe_Bruise	28
8.	Emaciated	44
9.	More-Than-One-Bruise	28
10.	DH2	114
11.	DF2	9
12.	Broken-Bone-Haemorage	4
13.	Condemned-DOA	1
14.	Condemned-Emaciated	2
15.	Condemned-Dead-Dying	2

Grade_Code

1.	AP5	4,285
2.	AP3	9,309
3.	AP4	3,048
4.	other	499
5.	AD	1,496
6.	AF1	322
7.	AP2	1,737
8.	AP1	155
9.	unknown	51
10.	TBR	28
11.	AM	424
12.	AF2	100

Analysis with 1Rw produced the following results :
Bruised :
default class : ‘none’ with 19,910/21,448
best hypothesis : ‘farm’ with 93% accuracy.

Bruise_Code :
default class : ‘none’ with 19,910/21,448
best hypothesis : ‘farm’ with 93% accuracy.

Grade_Code :
default class : ‘AP3’ with 9,309/21,448
best hypothesis : ‘hcw’ with 88% accuracy.

Data analysis procedure	**Results**
A new attribute, ‘days_held’, was created to combine attributes ‘month’, ‘day’, ‘sltr_month’, and ‘sltr_day’ to give a value for how many days held in the slaughter house. An attribute ‘farm_size’ was created as the number of deer slaughtered from each farm.	

Six instances were removed; one had errors in the date, and five contained many missing values. 1. C4.5 was run using 'Bruised' as a class.	1. The simplified decision tree many errors
Bruising was then made into a new attribute (Bruised). 2. C4.5 was run using 'Bruised' as a class. dominant attribute : origin	2. The simplified decision tree had less errors dominant attributes : origin
Bruising was then made into a new attribute (Bruise_Code) with 15 classes. 3. C4.5 was run using 'Bruising2' as a class.	3. The simplified decision tree had more errors
4. C4.5 was run using 'Grade_Code' as a class with all attributes.	4. The simplified decision tree was : size : 434 (nodes) with 32.3% errors dominant attributes : 'bruise_code' 'sex' 'gr' 'other_damage_code' 'distance'
5. C4.5 was run using 'Grade_Code' as a class with : 'sex' 'fat' 'days_held' 'distance_course' 'farmsize'.	5. This simplified decision tree was : *size : 46 with 41.9% errors* dominant attributes 'sex' 'fat'
6. C4.5 was then run with 'Grade_Code' as the class with attributes : 'Bruise_Code' 'Other_damage_code' 'fat' 'days_held'	6. The simplified decision tree was : size : 46 with 43.7% errors dominant attributes : 'Bruise_Code' 'other_damage_code'

Rules were previously made by Jenny Jago for classifying the animals by 'grade'. This was done using the attributes 'gr' (fat) and 'hcw'.	
7. C4.5 was used with 'Grade' as the class with attributes 'gr' and'hcw' to find an indication of how to classify grade.	7. The simplified tree indicated divisions between the classes at similar values of 'gr' and 'hcw' as the hand classified divisions.

The best result (no. 5) is presented at the end of this analyses.

13.3 Discussion of results

Several attributes were large enumerations, ie. they divide the data from one attribute into several values (eg. attribute 'farm' has 501 possible values). For machine learning this seemed be of little value since as farms usually submit only a single lot to the slaughter-house.

Farm was the attribute most influencing level of bruising, with the other attributes contributing minimally to this class. Sex and Bruise_Code had the most effect on grade. A decision tree for classifying fat ('gr') was found to be very similar to the classification techniques done by Jenny Jago.

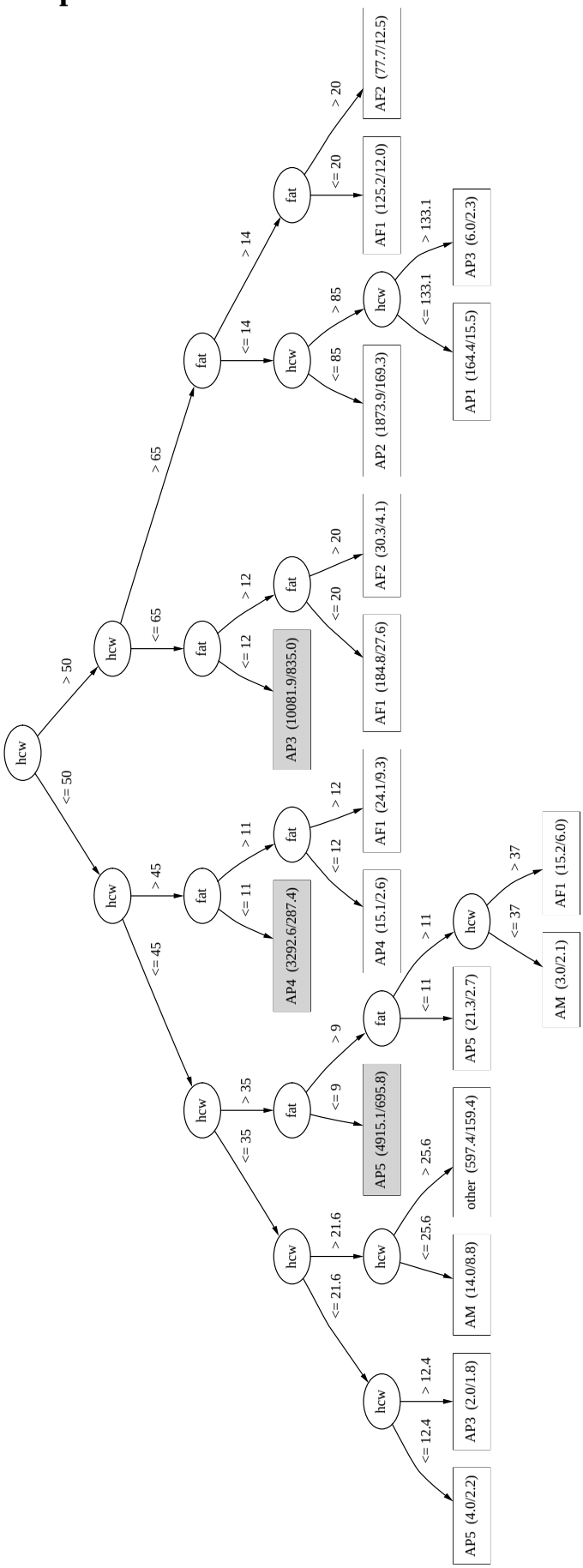
13.4 Conclusion

This dataset appeared to be quite suitable to machine learning since it has a discrete class (bruised or not bruised) and several other discrete attributes that could be used in a decision tree to describe the classes. However, none of the attributes that were previously found to contribute to bruising were suggested by the machine learning outputs. The format of this dataset allows it to be easily used by machine learning, but it didn't produce any valuable outcomes.

Follow up notes

Jenny Jago commented on the results. Her analysis of the results were that machine learning was able to detect the influence of farm of origin as the main contributing factor to bruising is not very important because generally, only one lot of animals came from each farm. Also, relationships between bruising and gr (fat level), sex of animal, and distance travelled did not show any clear relationships with machine learning, but they were found to be strongly correlated when using statistical techniques. Jenny was hoping that machine learning would be able to back up the relationships already found by statistics, but this was not done. Perhaps there is too much noise in the data.

Graph 1 : Result 5



14. WASP NESTS

Data source Andrew Wallace
Crop & Food Research
Private Bag 4704
Christchurch

Report date January, 1996

Assigned to Stephen Garner, Research programmer

Data location /home/ml/datasets/Summer1995/wasps/W188.arff.gz
/home/ml/datasets/Summer1995/wasps/W288.arff.gz
/home/ml/datasets/Summer1995/wasps/W89.arff.gz

14.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives

To record and compare aspects of nests of two wasp species (defined in terms of another experiment). This is to be able to improve the understanding of the wasp ecology as a contribution to biological control. The features recorded are mentioned below.

Summary of original research results

The results indicated that there was very little to distinguish between the nests of the different species with respect to nest sites and from spring to early summer in nest traffic, nest size and the number of combs. Some German wasp nests managed to survive the winter, but all common wasps had died by late June. Another interesting result indicated that common wasps tended to have nest entrances favouring the morning sun.

14.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective

To see by using the features recorded for each of the nests, a model could be induced to determine the wasp species that built the nest.

Dataset description
The dataset descriptions include instance, attribute, and class information. The original dataset was made up of three files. These were data from the 1988 period in the files W188.DAT and W288.DAT, and the data for the 1989 period in W89.DAT. The data in the W188 and W288 data files were combined to produce a new data file, W88, that had the same attributes as W89 in the same order. To do this, a derived attribute 'traffic_2' was created from the average of the existing attributes 'in' and 'out'. The final dataset is recorded below.

Instance information
A high number of instances contained missing values and noise. The factors present are only a representation of all the features required to describe

1988 dataset : 280 instances.
1989 dataset : 226 instances.

Attribute information
Number of attributes : 13 (12 actual attributes plus one predefined class).

W88
The attributes are :

1. Species (class)
   type : enumerated
   values : vulgaris, germanica
   missing : 0
2. Area
   type : integer
   values : 2.0 - 86.0
   missing : 5
3. Month
   type : enumerated
   values : -5 - 7
   missing : 17
4. Direction
   type : enumerated
   values : 1 - 9 (1 = NE, 2 = N, 3 = UP, 4 = NW, 5 = W, 6 = SW, 7 = S, 8 = SE, 9 = E)
   missing : 60
5. Traffic_2
   type : real
   values : 0.0 - 173.0
   missing : 64
6. Nest_site
   type : enumerated
   values : 1 - 10 (1 = soil, 2 = compost, 3 = wood, 4 = roots, 5 = rubbish, 6 = hanging, 7 = straw, 8 = debris, 9 = sleeper, 10 = other)
   missing : 9

7. Tunnel_length
   type : real
   values : 0.0 - 150.0
   missing : 64
8. Width
   type : real
   values : 3.0 - 70.0
   missing : 64
9. Height
   type : real
   values : 3.0 - 66.0
   missing : 65
10. Depth
   type : real
   values : 2.0 - 70.0
   missing : 65
11. Cavity
   type : enumerated
   values : 1 - 10
   missing : 18
12. Layers
   type : integer
   values : 1.0 - 25.0
   missing : 54
13. Castes
   type : enumerated
   values : 1, 2, 3 (1 = worker, 2 = male, 3 = queen)
   missing : 58

W89

The attributes are :

1. Species (class)
   type : enumerated
   values : vulgaris, germanica
   missing : 0
2. Area
   type : integer
   values : 1.0 - 88.0
   missing : 3
3. Month
   type : enumerated
   values : 1 - 9
   missing : 0
4. Direction
   type : enumerated
   values : 1 - 9 (1 = NE, 2 = N, 3 = UP, 4 = NW, 5 = W, 6 = SW, 7 = S,
   8 = SE, 9 = E)
   missing : 77

5. Traffic_2
   type : real
   values : 0.0 - 109.5
   missing : 69
6. Nest_site
   type : enumerated
   values : 1 - 10 (1 = soil, 2 = compost, 3 = wood, 4 = roots, 5 = rubbish, 6 = hanging, 7 = straw, 8 = debris, 9 = sleeper, 10 = other)
   missing : 69
7. Tunnel_length
   type : real
   values : 0.0 - 110.0
   missing : 93
8. Width
   type : real
   values : 1.0 - 120.0
   missing : 114
9. Height
   type : real
   values : 4.0 - 100.0
   missing : 114
10. Depth
    type : real
    values : 1.0 - 67.0
    missing : 114
11. Cavity
    type : enumerated
    values : 1 - 10
    missing : 108
12. Layers
    type : integer
    values : 1.0 - 25.0
    missing : 98
13. Castes
    type : enumerated
    values : 1, 2, 3 (1 = worker, 2 = male, 3 = queen)
    missing : 91

When experimenting with the data in machine learning schemes, a limited number of attributes were used. These were :

- Species - derived from Species
- Direction - derived from Direction
- Month - decode code to appropriate MONYY format, derived from Month
- Castes - derived from Castes
- Month_Code - eg Jan, derived from Month
- Year_Code - eg YY (88), derived from Month
- Nest_Site - derived from Nest_site
- Nest_Volume - derived from Width * Height * Depth

Class information

For both datasets Species is used as the class. Species has two class values.

1. vulgaris
2. germanica

The distribution of the classes are :

W88

1. vulgaris	198
2. germanica	82

W89

1. vulgaris	167
2. germanica	59

Analysis using 1Rw produced the following results :

W88
default class : ‘vulgaris’ with 198/280
best hypothesis : ‘nest_no’ with 98% accuracy.

W89
default class : ‘vulagaris’ with 167/226
best hypothesis : ‘month’ with 78% accuracy.

Data analysis procedure	**Results**
1. C4.5 was run on the 1988 data with ‘species’ as the class using all other attributes.	1. The simplified decision tree was size : 15 (nodes) with 13.2% errors dominant attributes were ‘month’ ‘castes’ ‘depth’
2. C4.5 was run on the 1989 data with ‘species’ as the class with all other attributes. The 1988 and 1989 data sets were combined	2. The simplified decision tree was size : 5 with 21.7% errors dominant attributes ‘month’ (Nov - May, Jun - Sep) ‘traffic_2’ (<=7 vs >7)
3. C4.5 was run with ‘species’ as the class.	3. The simplified decision tree was size : 21 with 16% errors dominant attributes were ‘month’ ‘height’ ‘depth’.

14.3 Discussion of results

The main attribute used to determine which species is time of year or month. Species*Vespula germanica* nests mainly over the winter, while nests of the species*Vespula vulgaris* are found mainly during the summer months.

14.4 Conclusion

The results indicate that the data taken for a different experiment are not compatible with the data required to run the machine learning experiments required here.

Reference : Donovan B.J., Howie A.M.E., Schroeder N.C., Wallac A.R. and Read P.E.C. (1992) “Comparative characteristics of nests of *Vespula germanica* (F.) and *Vespula vulgaris* (L.) (Hymenoptera : Vespinae) from Christchurce City, New Zealand” *NZ Jouranl of Zoology*, 1992, Vol 19: 61 - 71.

15. WHITE CLOVER PERSISTENCE TRIALS

Data source Ian Tarbotton
AgResearch
Whatawhata Research Centre
Hamilton

Report date January 1996

Assigned to Colleen Burrows, Research programmer

Data location /home/ml/datasets/Summer1995/pasture/REPORT/results1.arff
/home/ml/datasets/Summer1995/pasture/REPORT/results2.arff

15.1 Original research

The original research is discussed in terms of the objectives and subsequent results.

Objectives
To determine the mechanisms which influence the persistence of white clover populations in summer dry hill land. In particular reference to the consequence of a severe summer dry period in 1993/1994 and how it impacted on the performance of three white clover cultivars in an on-going experiment located at Whatawhata Research Centre.

Summary of original research results
Three key results were found; these included a) total white clover content was determined more by environmental differences than plant genetics, b) with time the proportion of resident genotypes contributing to the total population increased, c) resident genotypes appear well adapted to the experimental conditions.

15.2 Machine learning

Machine learning is discussed in terms of the machine learning objective, a description of the dataset, the analysis procedure followed, and subsequent results.

Objective
To predict the amount of White Clover in 1994 from the amount of White Clover and other species in the years 1991 to 1994 as well as information on the ‘strata’ where the White Clover was being grown.

Dataset description
The dataset descriptions include instance, attribute, and class information.

Instance information
The original data present had few errors and noise. The factors present are only a representation of all the features required to describe White Clover Persistence Trials.

Attribute information
Number of attributes in original dataset: 12 (11 actual attributes plus one predefined class), number of instances 14,400..

We tried to predict the amount of White Clover in 1994 from the amount of White Clover and other species in the years 1991 to 1994 as well as information on the strata where the White Clover was being grown. OtherGrasses-94 and Cocksfoot-93 and 94 seemed to influence the amount of White Clover the most, but there didn't seem to be one or two variables that consistently contributed to the class.

The attributes in the original dataset were :

1. Year
   type : enumerated
   values : 1991 - 1994
   missing : 0
2. Paddock
   type : enumerated
   values : 25, 42, 45
   missing : 0
3. Transect
   type : enumerated
   values : 1 - 30
   missing : 0
4. Plot
   type : enumerated
   values : 1 - 3
   missing : 0
5. Strata
   type : enumerated
   values : 1 - 7
   missing : 0
6. WhiteClover
   type : enumerated
   values : yes, no
   missing : 0
7. BareGround
   type : enumerated
   values : yes, no
   missing : 0
8. Cocksfoot
   type : enumerated
   values : yes, no
   missing : 0

9. OtherGrasses
   type : enumerated
   values : yes, no
   missing : 0
10. OtherLegumes
   type : enumerated
   values : yes, no
   missing : 0
11. RyeGrass
   type : enumerated
   values : yes, no
   missing : 0
12. Weeds
   type : enumerated
   values : yes, no
   missing : 0

Using the attribute editor, the data was then combined so that there was one instance for each strat-paddock-plot combination. The years were also combined so that there are attributes such as WhiteClover-91, WhiteClover-92, WhiteClover-93, etc, for all species. In each of these attributes, the value is a number which represents all of the 'yes' values for that strata-paddock-plot combination in the given species-year. This results in 63 instances with 31 attributes.

There are 32 attributes for the edited dataset. These are:

1. Strata
   type: enumerated
   values:1_OldCamp; 2_NewTrack; 3_Easy; 4_Moderate; 5_Steep; 6_OldEdge; 7_NewEdge
   missing: 0
2. Plot
   type : enumerated
   values : huia; prop; tahora
   missing: 0
3. Paddock
   type : enumerated
   values : 24; 42; 45
   missing: 0
4. WhiteClover-91
   type : Real
   values : 0.0 - 39.13
   missing: 0
5. BareGround-91
   type : Real
   values : 0.0 - 33.33
   missing: 0

6. Cocksfoot-91
   type : Real
   values : 0.0 - 50.0
   missing: 0
7. OtherGrasses-91
   type : Real
   values : 0.0 - 50.0
   missing: 0
8. OtherLegumes
   type : Real
   values : 0.0 - 20.0
   missing: 0
9. RyeGrass-91
   type : Real
   values : 0.0 - 57.14
   missing: 0
10. Weeds-91
   type : Real
   values : 0.0 - 40.0
   missing: 0
11. WhiteClover-92
   type : Real
   values : 0.0 - 50.0
   missing: 0
12. BareGround-92
   type : Real
   values : 0.0 - 10.71
   missing: 0
13. Cocksfoot-92
   type : Real
   values : 0.0 - 100.0
   missing: 0
14. OtherGrasses-92
   type : Real
   values : 0.0 - 42.48
   missing: 0
15. OtherLegumes-92
   type : Real
   values : 0.0 - 11.82
   missing: 0
16. RyeGrass-92
   type : Real
   values : 0.0 - 100.0
   missing: 0
17. Weeds-92
   type : Real
   values : 0.0 - 33.33
   missing: 0

18. WhiteClover-93
    type : Real
    values : 0.0 - 31.71
    missing: 0
19. BareGround-93
    type : Real
    values : 0.0 - 7.69
    missing: 0

20. Cocksfoot-93
    type : Real
    values : 0.0 - 58.33
    missing: 0
21. OtherGrasses-93
    type : Real
    values : 0.0 - 41.58
    missing: 0
22. OtherLegumes-93
    type : Real
    values : 0.0 - 19.05
    missing: 0
23. RyeGrass-93
    type : Real
    values : 0.0 - 50.0
    missing: 0
24. Weeds-93
    type : Real
    values : 0.0 - 40.0
    missing: 0
25. WhiteClover-94
    type : enumerated
    values : 0<=WhiteClover-94<8.8225; 8.225<=WhiteClover-94<17.645; 17.645<=WhiteClover-94<26.4675; 26.4675<=WhiteClover-94<35.29
    missing: 0
26. BackGround-94
    type : Real
    values : 0.0 - 14.29
    missing: 0
27. Cocksfoot-94
    type : Real
    values : 0.0 - 100.0
    missing: 0
28. OtherGrasses-94
    type : Real
    values : 0.0 - 50.0
    missing: 0

29. OtherLegumes-94
    type : Real
    values : 0.0 - 25.0
    missing: 0
30. RyeGrass-94
    type : Real
    values : 0.0 - 75.0
    missing: 0
31. Weeds-94
    type : Real
    values : 0.0 - 50.0
    missing: 0
32. Strata-Combined
    type : enumerated
    values : 1; 3; 4
    missing: 0

(this means that there are 31 actual attributes, one class attribute)

Original dataset : 14,400 instances, 12 attributes.
Final datasets : 63 instances, 32 attributes

Class information

The dataset contained two classes, which were ‘White-Clover-94’ and ‘Strata’ with the following values :

White-Clover-94

1. 0<=WhiteClover-94<8.8225
2. 8.8225<=WhiteClover-94<17.645
3. 17.645<=WhiteClover-94<26.4675
4. 26.4675<=WhiteClover-94<=35.29

Strata

1. 1_OldCamp
2. 2_NewTrack
3. 3_Easy
4. 4_Moderate
5. 5_Steep
6. 6_OldEdge
7. 7_NewEdge

The distribution of the class values are :

White-Clover-94

1.	0<=WhiteClover-94<8.8225	38
2.	8.8225<=WhiteClover-94<17.645	20
3.	17.645<=WhiteClover-94<26.4675	4
4.	26.4675<=WhiteClover-94<=35.29	1

Strata

1.	1_OldCamp	9
2.	2_NewTrack	9
3.	3_Easy	9
4.	4_Moderate	9
5.	5_Steep	9
6.	6_OldEdge	9
7.	7_NewEdge	9

Analysis with 1Rw produced the following results :

White-Clover-94
default class : '0<=WhiteClover-94<8.8225' with 38/63
best hypothesis : 'BareGround' with 73% accuracy

Strata
default class : '1_OldCamp' with 9/63
best hypothesis : 'RyeGrass' with 56% accuracy.

Data analysis procedure	**Results**
The data was combined where : • one instance was produced for each strata-paddock-plot • years for attributes such as WhiteClover## for all species are now present.	
1. C4.5 was run with WhiteClover-94 as a class with all other attributes. Ian Tarbotton suggested that some of the strata could be combined because of similarity. Strata [1, 2, 6, and 7] and [4 and 5] were combined.	1. The simplified decision tree was : *size 17 (nodes) 3.2% errors* dominant attributes were : 'OtherGrasses-94' 'Strata' 'Cocksfoot-94' 'Cocksfoot-93'
2. C4.5 was run with WhiteClover-94 as the class, with the new dataset formulated.	2. The simplified decision tree was : size 19 4.8% errors dominant attributes were : 'OtherGrasses-94' 'Cocksfoot-93' 'Cocksfoot-94'.
3. C4.5 was run with 'Strata' as the class and with all of the other attributes.	3. The simplified decision tree was : size 21 6.3% errors dominant atttributes were : 'RyeGrass-93'

‘Ryegrass-94’ ‘BareGround-94’	
4. C4.5 was run with ‘Plot’ as the class with all other variables as attributes. Since this data was in a time series format, the data was combined into 2-year windows. Attributes created were WhiteClover-1, WhiteClover-2, BareGround-1, BareGround-2, etc for all species and a difference attribute taking the difference between year 1 and year 2.	4. The simplified decision tree was : size 17 11.1% errors dominant attributes were : ‘WhiteClover-91’ ‘Cocksfoot-92’ ‘OtherGrasses-92’
5. C4.5 was run with WCDiffCLass (difference between year one and two WhiteClover) as the class with all new attributes as well as strata, plot, and paddock	5. The simplified decision tree was size 61 8.5% errors dominant attributes were : ‘WhiteClover-1’ ‘OtherLegumes-2’ ‘RyeGrass-2’.
6. C4.5 was then run without using ‘Strata’ as an attribute.	6. The simplified decision tree was : size 63 7.9% errors dominant attributes were : ‘WhiteClover-1’ ‘RyeGrass-2’ ‘OtherLegumes-2’ Strata does not seem to have much of an influence on the difference between clover growth from one year to the next.
7. C4.5 was run with ‘Strata’ as the class with all other attributes on separate datasets for each cultivar.	7. a) In the huia dataset, the simplified decision tree was size 13 1 error dominant attributes ‘OtherLegumes-92’ ‘Cocksfoot-94’ ‘BareGround-91’ b) In the tahora dataset, the simplified decision tree was size 13 0 errors dominant attributes were : ‘OtherLegumes-93’ ‘WhiteClover-93’

'BareGround-91'	c) In the 'Prop' dataset, the simplifed decision tree was size 13 1 error dominant attributes were : 'OtherGrasses-93' 'WhiteClover-91' 'BareGround-91'
8. C4.5 was run with 'Plot' as a class using all other attributes.	8. The simplified decision tree was size 21 2 errors dominant attributes were : 'WhiteClover-92' 'OtherGrasses-91' 'OtherLegumes-94' This tree seems to be able to divide plot 25 from the other two plots (42 and 45) which was where the differences were thought to be.

The results are presented at the end of this analyses.

15.3 Discussion of results

The data was not set up for machine learning, neither was a real question asked or distinct classes to use. Fairly small decision trees with few errors were found in predicting the amount of White Clover in either 1994 or year 2 from strata and the amount of other species in previous years. The use of the forced subsetting option decreased the size of the decision tree in the windowed data by grouping some of the strata together so that a separate node was not required for each. When 'strata' was not used with the time windowed data a tree was made that is of similar size and errors as the tree which included strata. This demonstrates that strata may be one attribute that can be used to classify, but it does not need to be used. The key finding was that when the seven 'strata' were used as class, the C4.5 tree did not always clump together 'strata' that were put together by the researcher; some 'strata' that were put together by the researcher were also clumped by the tree, but some were not.

15.4 Conclusion

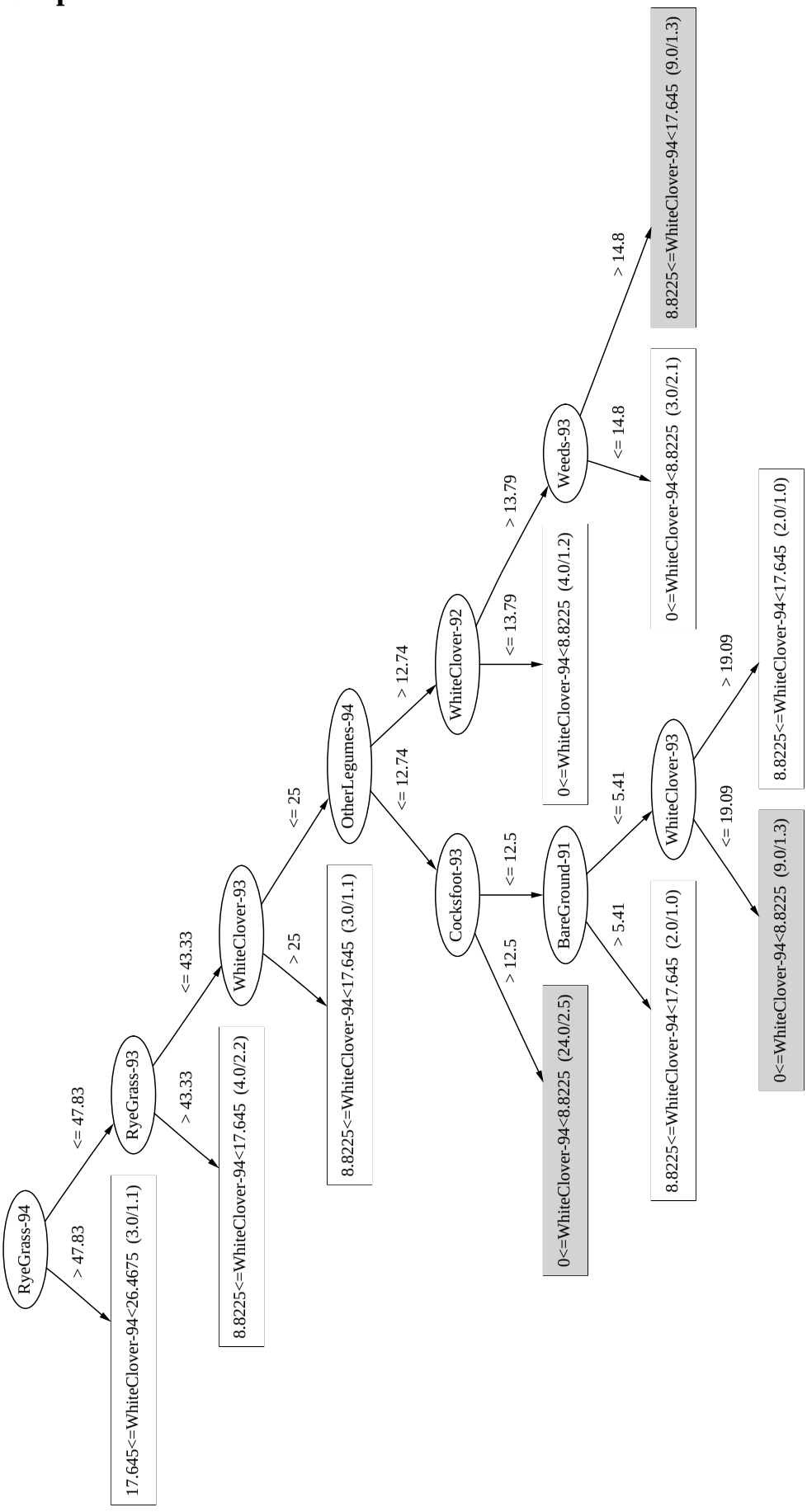
This dataset seemed to work well by finding which 'strata' are most closely related, but the question of which species influence White Clover growth was not well answered by machine learning.

Follow up notes

The results from which strata are clumped together is quite interesting. Brian de la Rue thinks that machine learning will work well for a first run through of a very large data set to be able to work through which relationships to statistically look at. Machine learning also has the strength that few assumptions need to be made before applying the machine learning. It also is quite good at giving a visual demonstration of some sort of tree. Disadvantages are that the decision trees are at first a little difficult to understand and so far there is no way to be able to give a measure of accuracy that would be acceptable for publication.

References : Sheath G.W., Wedderburn M.E., Tarbotton I.S. (1995) Persistence of Prop, Tahora and Huia White Colver in Summer Dry Hill Land*, A Contract Report to Agriseeds NZ Ltd.*

Graph 1 : Result 2



16. REFERENCES

Cameron-Jones, R.M., and Quinlan, J.R., 1993. Avoiding Pitfalls When Learning Recursive Theories. In: R. Bajcsy (Editor), Proc. IJCAI 93, Morgan Kaufmann, pp. 1050-1055.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D., 1988. AUTOCLASS: A Bayesian classification system. In: J. Laird (Editor), Proc. of the Fifth International Conference on Machine Learning. Ann Arbor, MI, Morgan Kaufmann, pp. 54-64.

Fisher, D., 1987. Knowledge Acquisition Via Incremental Conceptual Clustering. Machine Learning, 2: 139–172.

Gaines, B.R., 1991. The trade-off between knowledge and data in knowledge acquisition. In Piatetsky–Shapiro and Frawley, 1991, pp. 491–505.

Gennari, J.H., 1989. A survey of clustering methods. Technical Report 89-38, University of California, Dept. of Information and Computer Science, Irvine.

McQueen, R.J., Garner, S.R., Nevill-Manning, C.G., and Witten, I.G. Applying machine learning to agricultural data. Computers and Electronics in Agriculture (12) 1995, pp. 275-293.

Piatetsky–Shapiro, G. and Frawley, W.J. (Editors), 1991: Knowledge discovery in databases. AAAI Press, Menlo Park, CA., pp. 1-27.

Quinlan, J.R., 1990. Learning Logical Definitions from Relations. Machine Learning, 5: 239-266.

Quinlan, J.R., 1991. Determinate Literals in Inductive Logic Programming. In: J. Mylopoulos and R. Reiter (Editors), Proc. 12th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, pp. 746-750.

Quinlan, J.R., 1992. C4.5: Programs for Machine Learning. Morgan Kaufmann, pp. 1-16.

Quinlan, J.R., and Cameron-Jones, R.M., 1993. FOIL: a midterm report. In: P. Brazdil (Editor), Proc. European Conference on Machine Learning, Springer Verlag, pp. 3-20.