Working Paper Series
ISSN 1170-487X

# Analysis of Cow Culling Data with a Machine Learning Workbench

by Rhys E. DeWar[1] and
Robert J. McQueen[2]

Working Paper 95/1

January, 1995

© 1995 by Rhys E. DeWar & Robert J. McQueen
Department of Computer Science
University of Waikato
Private Bag 3105
Hamilton, New Zealand

[1] Department of Computer Science
[2] Department of Management Systems

# TABLE OF CONTENTS

**1. INTRODUCTION ..... 2**

**2. PROCEDURES USED TO PROCESS THE DATASETS ..... 3**

2.1 C4.5 ..... 3
*2.1.1 C4.5 Options Used ..... 3*
*2.1.2 Attributes Used ..... 4*

**3. ANALYSIS OF THE DATASETS FOR EACH PARTICIPANT ..... 5**

3.1 EXPECTED RESULTS ..... 5
3.2 BGPP ..... 6
3.3 BKQL ..... 7
3.4 BKYF ..... 8
3.5 BTWQ ..... 9
3.6 CDJX ..... 10
3.7 CHGX ..... 11
3.8 DYVT ..... 12
3.9 FMTP ..... 13
3.10 WCR ..... 14
3.11 WXV ..... 15
3.12 CONCLUSIONS ..... 16

**4. COMPARISON BETWEEN PARTICIPANTS ..... 17**

4.1 EXPECTED RESULTS ..... 17
4.2 FMTP DECISION TREE: ..... 18
4.3 FMTP VS. BGPP ..... 19
4.4 FMTP VS BKQL ..... 19
4.5 FMTP VS BKYF ..... 20
4.6 FMTP VS BTWQ ..... 20
4.7 FMTP VS CDJX ..... 21
4.8 FMTP VS CHGX ..... 21
4.9 FMTP VS DYVT ..... 22
4.10 FMTP VS WCR ..... 22
4.11 FMTP VS WXV ..... 23
4.12 CONCLUSIONS ..... 23

**5. ANALYSIS OF COMBINED PARTICIPANTS ..... 24**

5.1 EXPECTED RESULTS ..... 24
5.2 COMBINED PARTICIPANTS DECISION TREE ..... 25
5.3 COMBINED TREE VS. INDIVIDUAL PARTICIPANTS ..... 26
5.4 COMPARING FMTP TO THE COMBINED TREE ..... 27
5.5 CONCLUSIONS ..... 27

**6. CLASSIFYING ON PI CLASS ..... 28**

6.1 PI CLASS DISTRIBUTION ..... 28
*6.1.1 PI Class Tree ..... 29*
6.2 CONCLUSIONS ..... 30

**7. OVERALL CONCLUSIONS ..... 31**

7.1 PRE-PROCESSING OF THE DATA ..... 31
7.2 POSSIBLE INCONSISTENCY IN THE DATA ..... 31
7.3 THE PROBLEM OF SMALL DISJUNCTS ..... 31
7.4 CULLING/RETAINING DECISIONS ..... 32

**8. REFERENCES ..... 33**

# 1. Introduction

This report discusses the use of machine learning tools to examine datasets extracted from a database of dairy cows. The objective of the study was to investigate whether these machine learning tools could extract meaningful rules from the datasets, and in the process, understand more about how these tools manipulate data.

The preparation of the datasets for processing by the machine learning schemes was not a trivial activity. Two students spent a total of about 6 person months in 1993-1994 working with one of the 10 datasets in generating derived attributes that would produce usable results from the machine learning schemes. The requirement for this human-based pre-processing activity stemmed from the characteristic of the datasets that no one attribute adequately reflected the desired classification attribute (that a cow had been culled from the herd). In addition, other data transformation steps were required (age calculated from birth date relative to the year of the current record). Tools including a microcomputer spreadsheet, UNIX functions, and a database were used in various transformation steps. As a new derived attribute was developed, it was included in a test dataset and run through the machine learning schemes to determine its effect, modifications suggested, and the process iterated. Preliminary results from this single dataset testing have been previously reported [1,2,3,7].

In late 1994, a transformation script was developed and tested on the original test dataset (WXV) and the transformation results checked with the human-based result of the effort earlier in the year. When everything was correct, the script then processed the other nine datasets to produce datasets containing the derived attributes required in the format required by the machine learning schemes.

With the availability of the set of 10 datasets each containing the derived attributes which had been developed to investigate the "cow culled" classification attribute, a number of comparisons were possible. First, each of the 10 datasets were then processed individually through the machine learning scheme, and the decision trees that resulted were compared. Secondly, one dataset was used as a "training" dataset to develop rules, and then the rules were run against a second dataset and the accuracy calculated. This comparison was done on a number of pairs of datasets. Finally, the whole 10 datasets were combined into one large dataset (16615 records) and a decision tree generated.

Finally, some additional work was done investigating the accuracy of the breeding index (BI) as a predictor of classes of production from each cow, as measured by the production index (PI).

This work was undertaken by members of the machine learning research group, which is led by Prof. Ian Witten, and is based in the Department of Computer Science, University of Waikato. The work has been supported by a grant from the Foundation for Research, Science and Technology.

# 2. Procedures used to process the Datasets

The LIC cow-culling datasets were provided by the Livestock Improvement Corporation(LIC), an organisation which maintains a large relational database of dairy cow and sire production and breeding information.. The datasets provided were extracts from this large database, and incorporated data from the data for 10 Participant's dairy herds. Each 'Participant' is the owner of a farm who agreed to have production information of their herds recorded by LIC for research. The extracts covered the 6 years from 1987 to 1992, although some of these years are not present for some of the participants.

The raw datafiles contain the details of individual cows in the participant's herd in the given year, including the cow's identification, sire and dam information, progeny information, and production and parturition details.

These datasets were processed, including derivation of new attributes from attributes already in the data, and prepared for analysis by the schemes in the WEKA Machine Learning Workbench.

## 2.1 C4.5

Most of the analysis of the cow-culling datasets was done with C4.5 [5,6], a Machine Learning Scheme for the top-down induction of decision trees. C4.5 starts with the full dataset, decides which attribute contributes most information in deciding on the classification of all the cases, and produces branches on values of that attribute. Then, for each branch, the process is repeated, except instead of using the full dataset, only those instances which match the criteria of the branch are used. If all the remaining instances are the same class, then a leaf is created which contains that class.

### 2.1.1 C4.5 Options Used

The C4.5 program has many options which can be used to optimise performance on a given dataset. Several of these options were used in the analysis;

#### *2.1.1.1 Minimum Number of Objects per branch*

This option tells C4.5 not to create a new tree branch unless that branch contains greater than or equal to the specified number of instances. This is useful for preventing "overfitting" of the data when the number of instances remaining to be classified is small.

#### *2.1.1.2 Pruning Confidence Level*

C4.5 induces a decision tree, and then attempts to prune that tree by collapsing branches together. To do this, it uses a T-test to determine whether or not a branch can be pruned. With the cow-culling datasets, the set of animals with class "dep_farmer" is typically only 5% of the total number of instances. This leads to the problem where C4.5 treats this whole class as noise in the data, and prunes the whole tree down to one leaf node. The default pruning confidence level of C4.5 is 25%. By increasing this level, the amount of pruning performed on the tree can be reduced.

#### *2.1.1.3 Iterative Mode using windowing*

When using this option, C4.5 randomly splits the dataset into a training window and a test set. Instead of starting at the root of the tree with the whole dataset, C4.5 uses the training window to induce a tree, then tests it's accuracy on the test set. If the tree falls below C4.5's accuracy criterion, then it adds misclassified instances from the training set and induces a new tree, repeating this process until the tree meets the accuracy criterion. If there are strong patterns in the data, then C4.5 should be able to create an accurate, generalised tree with relatively few instances.

### 2.1.2 Attributes Used

In all cases, the attributes used in these runs were attributes derived from the original data. The reasons for using these attributes were;

- Many of the original attributes had to be converted into a form meaningful to the workbench schemes

- Some attributes in the original data were non-linear, and had to be combined to have meaning to the workbench schemes

- Many of the original attributes were irrelevant to the classification

# 3. Analysis of the Datasets for Each Participant

The datasets for all years were combined for each participant, in order to find patterns in culling decisions which persisted between years. These datasets were analysed using C4.5 to create decision trees and rulesets. All of these trees were produced using standard mode (i.e. tree induced from the whole dataset). Many were produced with the minimum number of objects per branch set to prevent overfitting.

The (pruned) decision tree for each participant is shown, along with a table showing the size and error rates of the pruned and unpruned trees, and a confusion matrix showing the numbers of instances classified in each class against their actual classes.

## 3.1 Expected Results

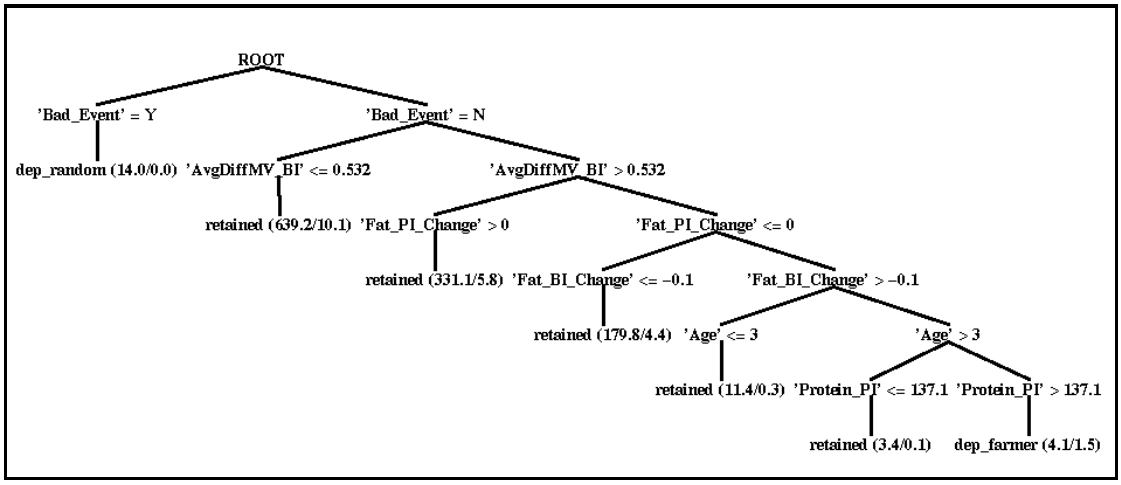
With the combined datasets, we expected decision trees which classified animals with good PI and/or BI values to be retained and animals with poor PI's and BI's to be culled. We also expected animals under 2 years old to be mainly retained (these animals have not yet begun production), and for 2 year olds to have different culling classifications (as many 2 year olds are culled if their first production run is poor).

The words 'good' and 'bad' used to describe the trees is a subjective term to describe how closely it matched our expectations based on our information about cow-culling.

## 3.2 BGPP

C4.5 Options :

Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 15 | 19(1.6%) | 13 | 19(1.6%) | 1.9% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 3 | 19 | | **culled** | **Actual** |
| 4 | 1147 | | **retained** | |
| | | 14 | **died** | **Class** |

Summary :

If a bad event has happened to the cow then dep_random
Otherwise, if Milk Volume BI is below the herd average by more than 0.532 then retained
Otherwise, if Fat PI Change is greater than 0 then retained
Otherwise, if Fat BI has decreased by more than 0.1 then retained
Otherwise, if Age is 3 or less then retained
Otherwise if Protein PI is 137.1 or less then retained
Otherwise culled

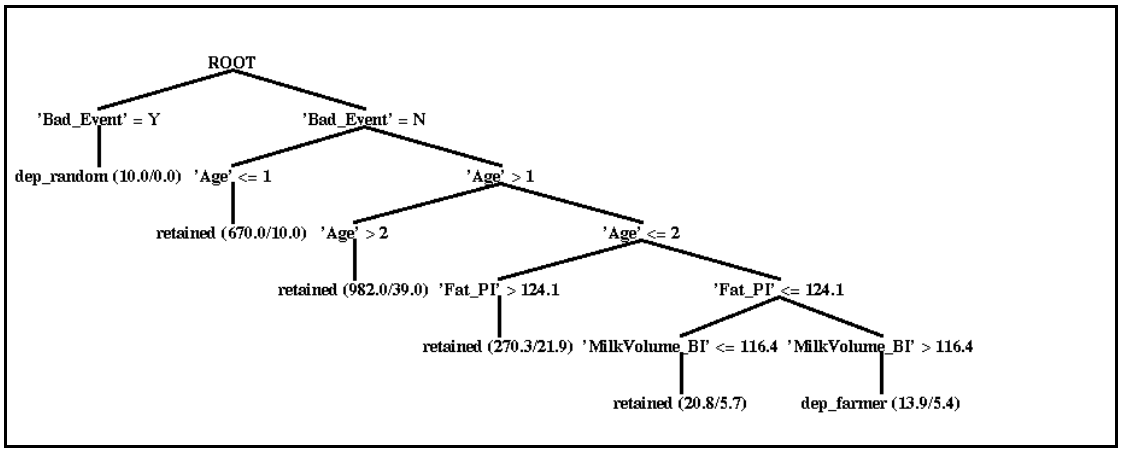
This tree seems reasonable. Retaining animals with below average milk volume BI can possibly be explained by the Dairy Co. penalties for volume. However, retaining animals whose Fat BI has gone down and culling animals with high Protein PI's seem illogical. The latter may be caused by there being only a small number of cases left to classify at that branch.

Note : This tree has no minimum number of objects per branch, because C4.5 tended to over-generalise on this data, creating a tree with only 2 leaves - one dep_random and the other retained.

## 3.3 BKQL

C4.5 Options :

Minimum of 8 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 11 | 79(4.0%) | 11 | 79(4.0%) | 4.3% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 8 | 77 | | **culled** | **Actual** |
| 2 | 1870 | | **retained** | |
| | | 10 | **died** | **Class** |

Summary :

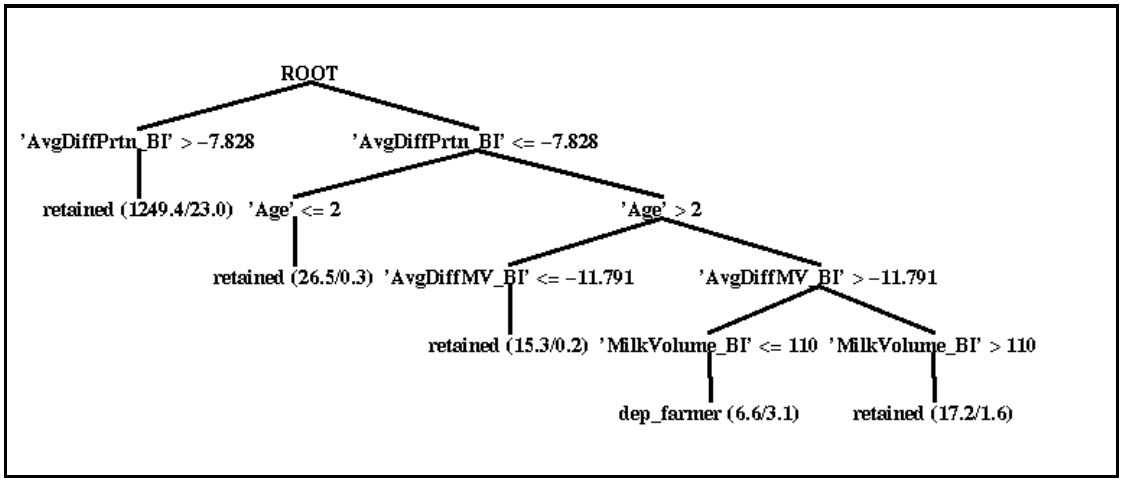
If a bad event has happened to the cow then dep_random
Otherwise, if age is 1 or less then retained
Otherwise, if age is more than 2 then retained (leaving only 2 year olds)
Otherwise, if Fat PI is more than 124.1 then retained
Otherwise, if Milk Volume BI is 116.4 or less then retained
Otherwise culled.

Good, ties in with what we were told about culling decisions for 2 year olds. Seems to be that high fat production is good, but if high milk production is undesirable, which seems to tie in with what we were told about penalty rates for too much volume.

## 3.4 BKYF

C4.5 Options :

Minimum of 5 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 31 | 20(1.5%) | 9 | 25(1.9%) | 2.1% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 4 | 24 | | **culled** | **Actual** |
| 1 | 1286 | | **retained** | |
| | | | **died** | **Class** |

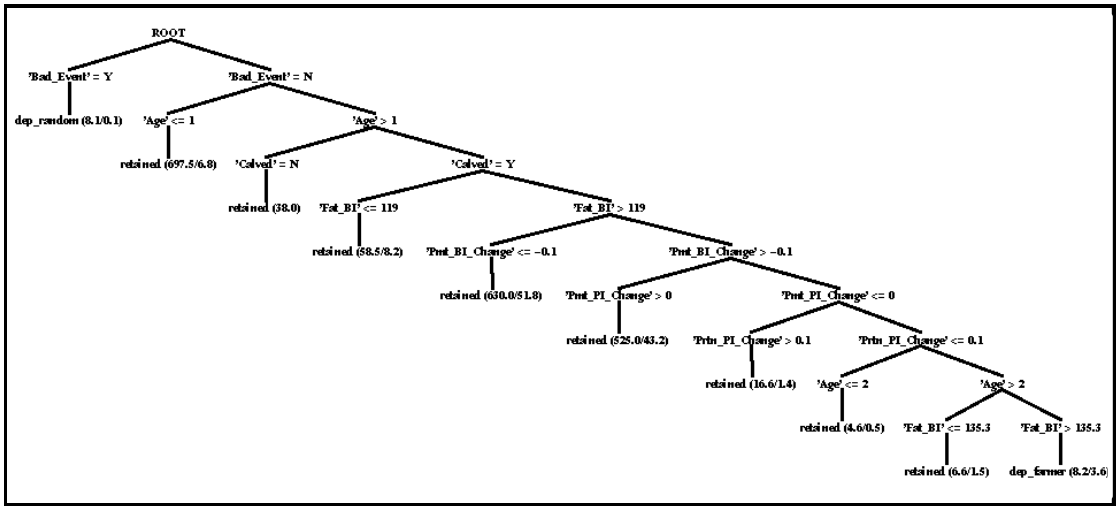
Summary :
If Protein BI is not more than 7.828 below the herd average then retained
Otherwise, if age is 2 or less then retained (leaving only 3+ year olds)
Otherwise, if Milk Volume BI is 11.79 or more below the herd average then retained
Otherwise, if Milk Volume BI is 110 or less then culled
Otherwise retained

Good; the bottom branches seem to define an acceptable milk volume level, between producing too little and producing too much.

## 3.5 BTWQ

C4.5 Options :

Minimum of 4 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 19 | 115(5.8%) | 3 | 118(5.9%) | 6.0% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 4 | 114 | | **culled** | **Actual** |
| 1 | 1866 | | **retained** | |
| | | 8 | **died** | **Class** |

Summary :

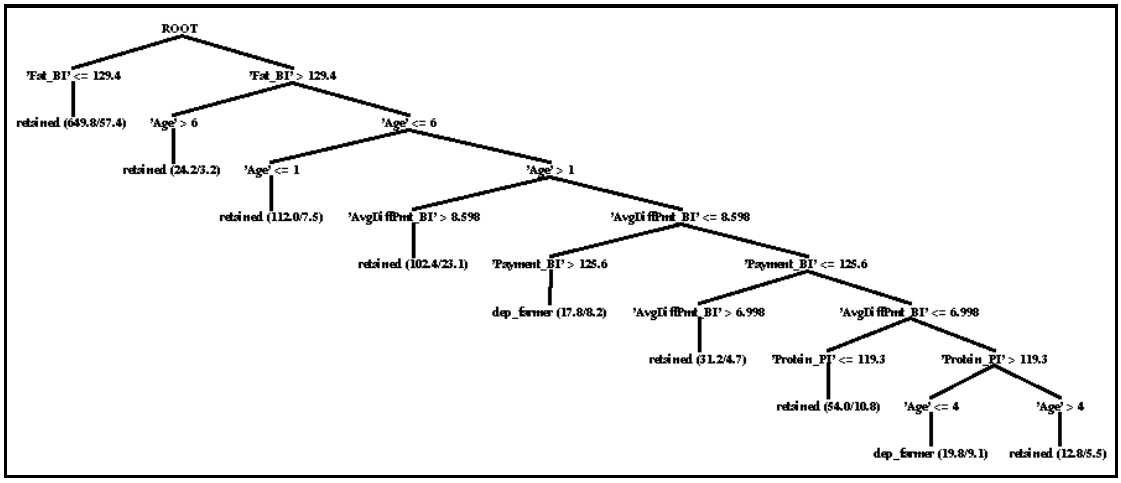
If a bad event has happened to the cow then dep_random
Otherwise, if age is 1 or less then retained
Otherwise, if the animal has not calved then retained
Otherwise, if Fat BI is less than or equal to 119 then retained
Otherwise, if Payment BI has decreased by 0.1 or more then retained
Otherwise, if Payment PI has increased then retained
Otherwise, if Protein PI has increased by more than 0.1 then retained
Otherwise, if age is 2 or less then retained
Otherwise, if Fat BI is 135.3 or less then retained
Otherwise culled

This tree contains some expected splits, but also some unexpected (and seemingly erroneous) splits, possibly indicating that the patterns in this data aren't well defined enough for C4.5 to find with the given attributes. In particular, retaining animals whose Payment BI has decreased seems illogical.

## 3.6 CDJX

C4.5 Options :

Minimum of 11 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 41 | 103(10.1%) | 17 | 112(10.9%) | 12.7% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 21 | 105 | | **culled** | **Actual** |
| 3 | 891 | | **retained** | |
| | 4 | | **died** | **Class** |

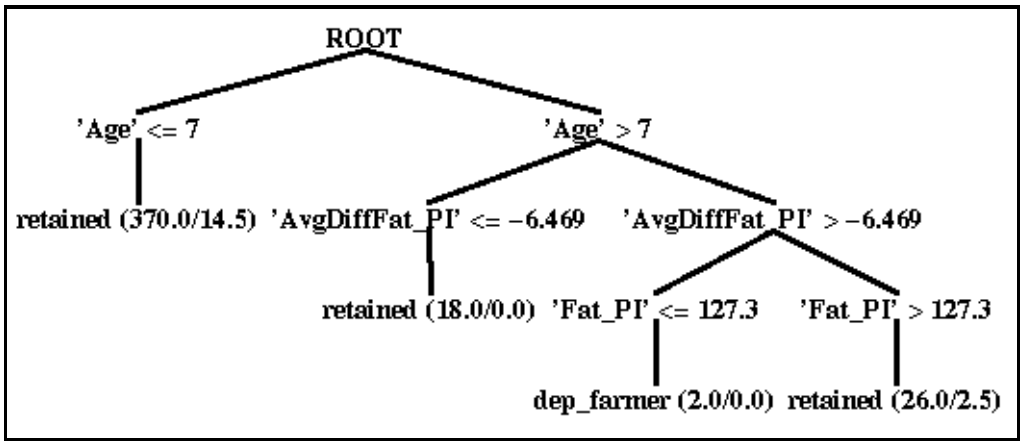
Summary :
If Fat BI is 129.4 or less then retained
Otherwise, if age is more than 6 then retained
Otherwise, if age is 1 or less then retained (leaving 2 to 6 year olds)
Otherwise, if Payment BI is 8.598 or more above average then retained
Otherwise, if Payment BI is more than 125.6 then culled
Otherwise, if Payment BI is more than 6.998 above average then retained
Otherwise, if Protein PI is 119.3 or less then retained
Otherwise, if age is 4 or less then culled
Otherwise retained

This tree seems logical based on what we have been told. Only the split on Protein PI (retaining animals below a certain PI) is questionable. This split may be due to most instances at that branch having an average or better Protein PI.

# 3.7 CHGX

C4.5 Options :

Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 9 | 16(3.8%) | 7 | 16(3.8%) | 4.1% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 2 | 16 | | **culled** | **Actual** |
| | 398 | | **retained** | |
| | | | **died** | **Class** |

Summary :

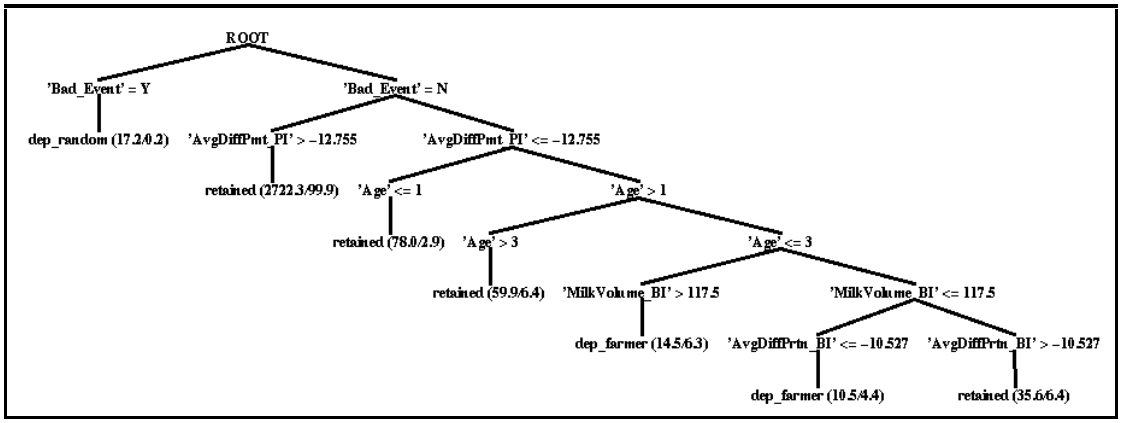
If age is 7 or less then retained
Otherwise, if Fat PI is 6.469 or more below average then retained
Otherwise, if Fat PI is 127.3 or less then culled
Otherwise retained

This tree seems good, although retaining all animals under seven years seems strange (14 cases incorrect at that branch shows that animals under 7 do get culled). Other than that, this participant seems to depend on Fat PI, culling those producing above or below a certain range.

# 3.8 DYVT

C4.5 Options :

Minimum of 8 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 13 | 121(4.1%) | 13 | 121(4.1%) | 4.4% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 14 | 116 | | **culled** | **Actual** |
| 5 | 2786 | | **retained** | |
| | | 17 | **died** | **Class** |

Summary :

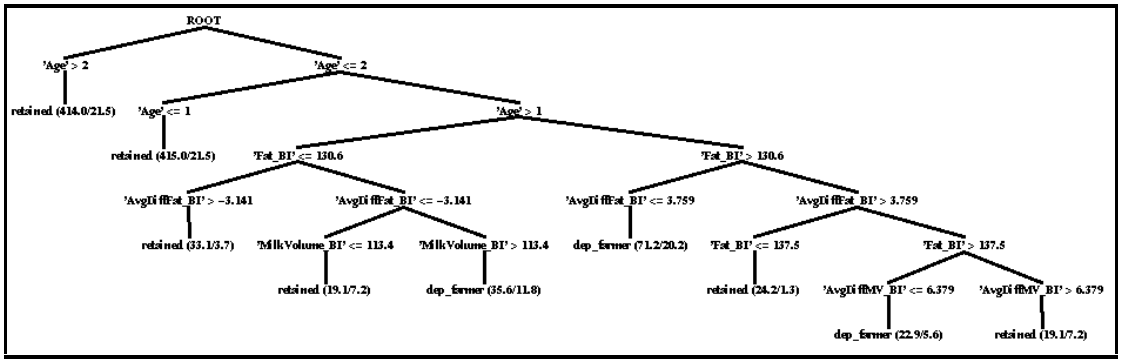
If Payment PI is better than 12.755 below average, then retained
Otherwise, if age is 1 or less, then retained
Otherwise if age is more than 3, then retained (leaving only the 2-3 year olds!)
If Milk Volume BI > 117.5 then culled
Otherwise, if Protein BI is worse than 10.527 below the herd average, then culled
Otherwise retained.

A very good tree, consistent with what we've been told.

## 3.9 FMTP

C4.5 Options :

Minimum of 10 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 25 | 79(7.5%) | 17 | 77(7.3%) | 9.5% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 89 | 58 | | **culled** | **Actual** |
| 12 | 888 | | **retained** | |
| 1 | 6 | | **died** | **Class** |

Summary :

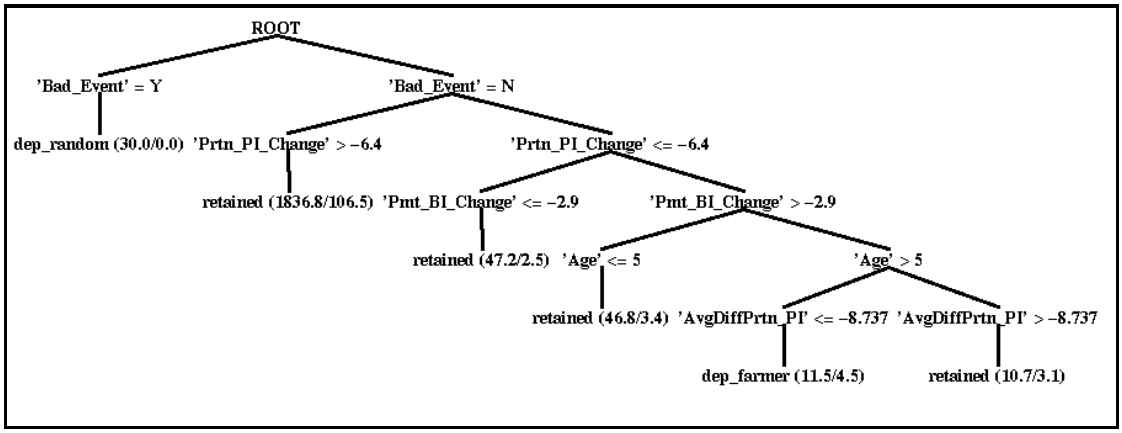
If age is more than 2, then retained
Otherwise, if age is 1 or less, then retained (leaving only the 2 year olds!)
If Fat BI is less than 130.6 then
  If Fat BI is better than 3.141 below the herd average then retained
  Otherwise if Milk Volume BI is greater than 113.4 then culled
  Otherwise retained
Otherwise if Fat BI is not more than 3.759 above the herd average then culled
Otherwise if Fat BI is less than 137.5 then retained
Otherwise if Milk Volume BI is more than 6.379 above the herd average then retained
Otherwise culled.

Very good. Better classification of culled instances than for the other participants.

# 3.10 WCR

C4.5 Options :

Minimum of 7 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 13 | 115(5.8%) | 11 | 115(5.8%) | 6.1% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 7 | 114 | | **culled** | **Actual** |
| 1 | 1831 | | **retained** | |
| | | 30 | **died** | **Class** |

Summary :

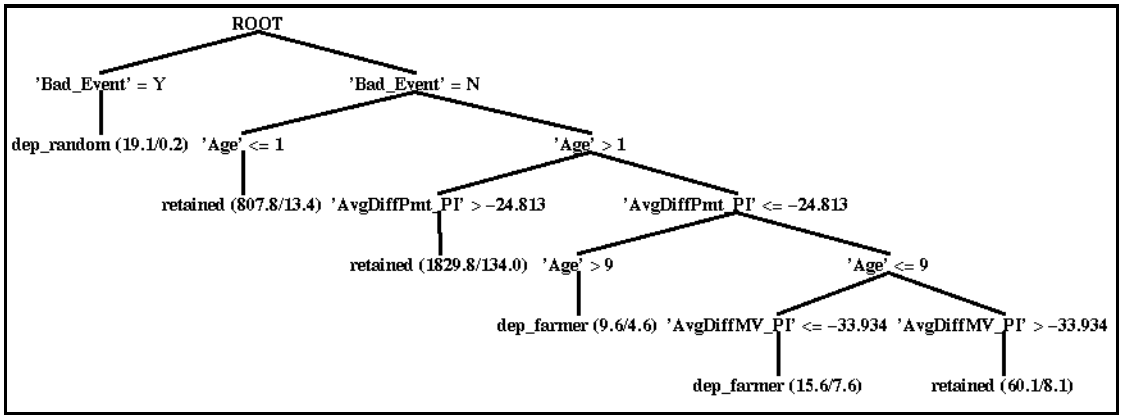
If a bad event happened to the cow then dep_random
Otherwise, if Protein PI has changed by more than -6.4 then retained
Otherwise, if Payment BI has decreased by more than 2.9 then retained
Otherwise, if age is 5 or less then retained
Otherwise, if Protein PI is 8.737 or more below average then culled
Otherwise retained

This tree seems quite good, apart from the branch retaining animals whose Payment BI has decreased by more than 2.9. This would seem to be a criterion for culling the animal rather than retaining it.

# 3.11 WXV

C4.5 Options :

Minimum of 8 Objects per Branch
Pruning Confidence level 99%




| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 13 | 159(5.8%) | 11 | 159(5.8%) | 6.1% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 13 | 155 | | **culled** | **Actual** |
| 4 | 2551 | | **retained** | |
| | | 19 | **died** | **Class** |

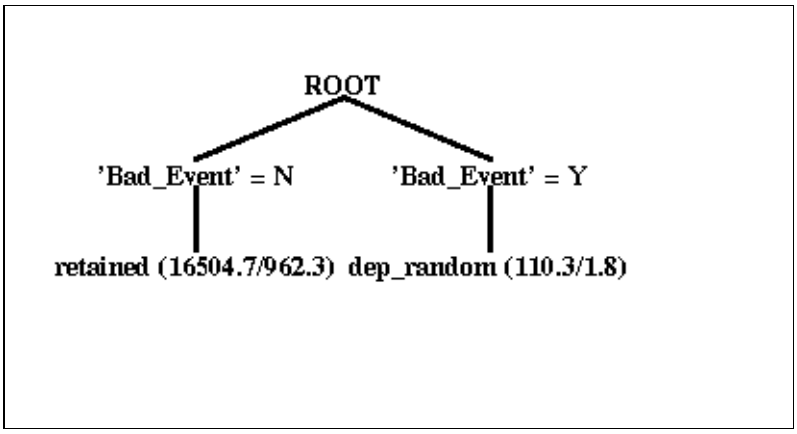
Summary :

If age is 1 or less then retained
Otherwise if Payment BI is better then 24.813 below the herd average then retained
Otherwise if age is more than 9 then culled
Otherwise if Milk Volume PI is below the herd average by 33.934 or more then culled
Otherwise retained

A very good tree, consistent with what we have been told by LIC.

## 3.12 Conclusions

All the trees produced classify the retained instances very well, but are not so good on the culled animals. The tree with the best performance for classifying culled instances is the tree for participant FMTP, with 89 instances correctly classified, but 58 misclassified as retained. This is caused by the culled instances being only a small percentage of the total number of instances - the trees are all "accurate" to around 5%, but the 5% error is almost all culled instances. In fact, for almost all the participants, the pruning confidence level had to be increased to avoid C4.5 producing a tree with only 2 leaves which did not classify any of the retained instances correctly, e.g.




However, almost all the trees produced seem to be logical, in terms of what we were expecting them to look like, i.e., that animals with high Fat, Protein and/or Payment PI/BI's were likely to be retained, and those with sub-standard PI/BI's were likely to be culled.

An interesting branch which appeared in the trees for several participants was culling animals with a high Milk Volume PI or BI. This suggests that very high milk production is undesirable, which may be based on Dairy Co. penalty rates for volume.

These trees are all based on one participant, and hence reflect the patterns within that herd, indicating culling/retaining rules used by that specific participant, rather than general culling/retaining rules.

# 4. Comparison Between Participants

In order to find how similar the culling (or retaining) rules used by the different participants were, a decision tree was induced from the dataset for one participant, and used to classify the datasets for the other participants.

Participant FMTP was chosen because the decision tree generated by FMTP had the best performance for culled animals. An extra branch was added to the tree to classify the animals which had died.

The FMTP decision tree is shown, along with it's confusion matrix and a summary. Then, for each other participant, an error table and confusion matrix are shown for how the FMTP tree performed on the dataset for that participant.

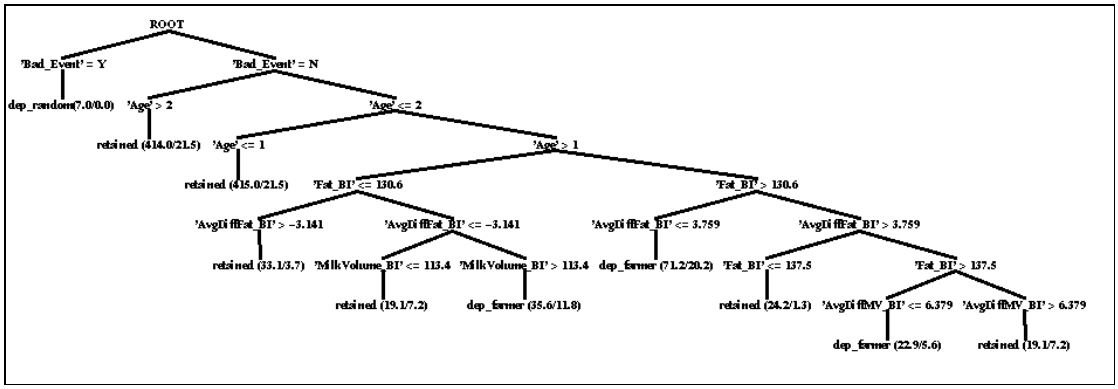
## 4.1 Expected Results

FMTP has the best performance on the culled instances, so we were interested to see if it performed well on the other participants - maybe the culling rules were similar between the participants, but more apparent in the FMTP dataset. However, because these participants may have different retaining/culling preferences, and their herds may have differing characteristics, very high performance was not expected.

## 4.2 FMTP Decision Tree:

C4.5 Options:

Trees evaluated on unseen cases
Minimum of 10 Objects per Branch
Pruning confidence level 99%




Evaluation on training data (1054 items):

| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 27 | 72(6.8%) | 19 | 70(6.6%) | 6.9% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 89 | 58 | | **culled** | **Actual** |
| 12 | 888 | | **retained** | |
| | | 7 | **died** | **Class** |

Summary

This decision tree is the best of all the trees for the participants at classifying the culled instances (89/147 = 61%). Only 12 (1.3%) of the retained instances are misclassified, and (with the 'extra' branch added), all the died instances are correctly classified.

## 4.3 FMTP vs. BGPP

Evaluation on test data (1183 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 103 | 8.7 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| | 22 | | **culled** | **Actual** |
| 81 | 1066 | | **retained** | |
| | | 14 | **died** | **Class** |

Summary
The entire culled class have been misclassified as retained, while 81 retained instances have been classified as culled. This indicates different culling methods or different herd characteristics between the two participants.

## 4.4 FMTP VS BKQL

Evaluation on test data (1967 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 206 | 10.5 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 10 | 75 | | **culled** | **Actual** |
| 131 | 1741 | | **retained** | |
| | | 10 | **died** | **Class** |

Summary
The FMTP tree classifies 10 of the 85 culled animals correctly, (better than the BKQL tree, which gets only 8 correct), but classifies 131 retained instances as culled (c.f. 2 for the BKQL tree). Again, this could indicate either differing culling methods or differing herd characteristics.

## 4.5 FMTP VS BKYF

Evaluation on test data (1315 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 109 | 8.3 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| | 28 | | **culled** | **Actual** |
| 81 | 1206 | | **retained** | |
| | | | **died** | **Class** |

Summary

Similar to the runs against the two above participants, all the culled instances have been classed as retained, while 81 retained instances have been classed as culled.

## 4.6 FMTP VS BTWQ

Evaluation on test data (1993 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 244 | 12.2 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| | 118 | | **culled** | **Actual** |
| 126 | 1741 | | **retained** | |
| | | 8 | **died** | **Class** |

Summary

This is similar to FMTP vs. BKYF.

## 4.7 FMTP VS CDJX

Evaluation on test data (1024 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 127 | 12.4 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 5 | 121 | | **culled** | **Actual** |
| 6 | 888 | | **retained** | |
| | | 4 | **died** | **Class** |

Summary
Similar to FMTP vs. BKYF, above. Performance in all classes is similar to the CDJX tree.

## 4.8 FMTP VS CHGX

Evaluation on test data (416 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 34 | 8.2 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| | 18 | | **culled** | **Actual** |
| 16 | 382 | | **retained** | |
| | | | **died** | **Class** |

Summary
Similar to FMTP vs. BKYF, above. The CHGX tree only got 2 culled instances correct.

## 4.9 FMTP VS DYVT

Evaluation on test data (2938 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 360 | 12.3 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 23 | 107 | | **culled** | **Actual** |
| 253 | 2538 | | **retained** | |
| | | 17 | **died** | **Class** |

Summary

Similar to FMTP vs. BKQL, above. 23 of the 130 culled instances have been classified correctly (c.f. 14 for the DYVT tree), but 253 of the retained instances have been classed as culled (c.f. 5 for the DYVT tree).

## 4.10 FMTP VS WCR

Evaluation on test data (1983 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 209 | 10.5 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 8 | 113 | | **culled** | **Actual** |
| 96 | 1736 | | **retained** | |
| | | 30 | **died** | **Class** |

Summary

Similar to FMTP vs. BKQL, above. 8 out of 121 culled instances have been correctly classified (c.f. 7 for the WCR tree), but 96 retained instances have been classified as culled (c.f. 1 for the WCR tree).

## 4.11 FMTP VS WXV

Evaluation on test data (1054 items):

| Error | |
|---|---|
| **No. of Errors** | **%age Error** |
| 239 | 8.7 |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 19 | 149 | | **culled** | **Actual** |
| 90 | 2465 | | **retained** | |
| | | 19 | **died** | **Class** |

Summary

Similar to FMTP vs. BKQL, above. 19 out of 168 culled instances have been correctly classified (c.f. 13 for the WXV tree), but 90 retained instances have been classified as culled (c.f. 4 for the WXV tree).

## 4.12 Conclusions

When run against the other participants, the FMTP generally classified the culled instances with the same or better performance as the tree for that particular participant, but classified a much higher number of retained instances as culled that the tree for that participant. This indicates, as expected, that there are significant differences in the culling patterns used by each participant. FMTP would cull many animals which the other participants would retain, indicating quite different culling/retaining rules.

Also see section 5.4 "**Comparing FMTP to the Combined Tree**" below.

# 5. Analysis of Combined Participants

In order to generate a general set of culling rules, we combined the datasets for all the participants into a combined dataset. This would generate a decision tree based on the general culling patterns across all participants.

Then, running the combined tree against the individual participant datasets would give an idea of the performance of this 'general' tree, and also the degree of influence that participant might have had on the tree.

## 5.1 Expected Results

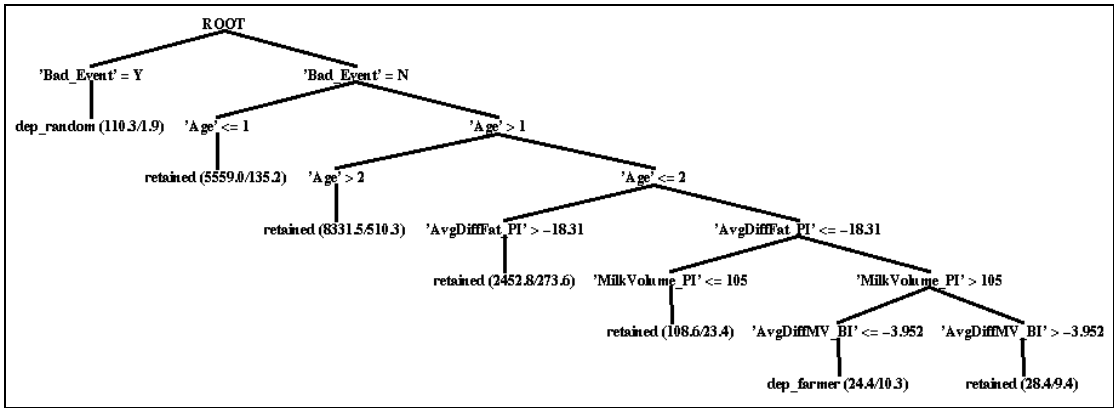
With the combined tree, we expected similar to the individual participants, good performance on the retained animals with poor performance on the culled instances. Because the tree was created from all the participant's datasets, it should produce branches which reflect the similarities in culling patterns between the participants.

## 5.2 Combined Participants Decision Tree

C4.5 Options

Minimum of 16 objects per branch

Pruning Confidence level 70%




Evaluation on training data (16615 items):

| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 35 | 942(5.7%) | 17 | 952(5.7%) | 5.8% |

| Classified as... | | | | |
|---|---|---|---|---|
| **culled** | **retained** | **died** | | |
| 14 | 949 | | **culled** | **Actual** |
| 3 | 15540 | | **retained** | |
| | | 109 | **died** | **Class** |

Summary

If a bad event has happened to the cow then died
Otherwise, if age is 1 or less then retained
Otherwise, if age is greater than 2 then retained (leaving only 2 year olds!)
Otherwise, if Fat PI is not more than 18.31 below the herd average then retained
Otherwise, if Milk Volume PI is less than 105 then retained
Otherwise, if Milk Volume BI is 3.952 above the herd average or less, then culled
Otherwise retained

As with the runs against the individual participants, all the died instances have been correctly classified, 99.9% of the retained instances have been correctly classified, but only 1.5% of the culled instances have been correctly classified.

## 5.3 Combined Tree vs. Individual Participants

| Classified as... | | | |
|---|---|---|---|
| **culled** | **retained** | **died** | |
| **BGPP** | | | |
| | 22 | | **culled** |
| | 1147 | | **retained** |
| | | 14 | **died** |
| **BKQL** | | | |
| 2 | 83 | | **culled** |
| | 1872 | | **retained** |
| | | 10 | **died** |
| **BKYF** | | | |
| | 28 | | **culled** |
| | 1287 | | **retained** |
| | | | **died** |
| **BTWQ** | | | |
| | 118 | | **culled** |
| 1 | 1866 | | **retained** |
| | | 8 | **died** |
| **CDJX** | | | |
| | 126 | | **culled** |
| | 894 | | **retained** |
| | | 4 | **died** |
| **CHGX** | | | |
| | 18 | | **culled** |
| 1 | 397 | | **retained** |
| | | | **died** |
| **DYVT** | | | |
| 6 | 124 | | **culled** |
| | 2791 | | **retained** |
| | | 17 | **died** |
| **FMTP** | | | |
| 5 | 142 | | **culled** |
| | 900 | | **retained** |
| | | 7 | **died** |
| **WCR** | | | |
| 1 | 120 | | **culled** |
| | 1832 | | **retained** |
| | | 30 | **died** |
| **WXV** | | | |
| | 168 | | **culled** |
| 1 | 2554 | | **retained** |
| | | 19 | **died** |

## 5.4 Comparing FMTP to the Combined Tree

The FMTP and combined trees are fairly similar, especially near the root. In order to find out how similar these trees are, NVOL [4] was used to compare the hyperspaces described by each tree (after serialising them into ripple-down rules). The resultant correlation matrix from running NVOL is;

| | | FMTP Rules | | |
|---|---|---|---|---|
| | | Died | Retained | Culled |
| Combined Rules | Died | 100.00% | | |
| | Retained | | 84.45% | |
| | Culled | | 0.07% | |

This indicates 100% correlation between classification of Died animals (expected, because each tree has only one branch for classifying this class on the same criterion), and a very strong correlation on the retained instances (again expected). However, there is no correlation between the classification of culled instances between the two rulesets. This indicates that the regions defined for culling animals in each ruleset only overlap an insignificant amount or not at all, and therefore that the rules for culling animals in each ruleset are very different. The combined tree succeeded in classifying 5 culled instances from the FMTP dataset, so there is a small amount of overlap.

A small amount of overlap is shown between the FMTP rules for the retained class and the combined rules for the culled class, meaning that a small number of instances which FMTP would retain would be culled by the combined ruleset. However, none of the FMTP instances lie in this region, as the combined tree did not misclassify any retained instances from the FMTP dataset.

## 5.5 Conclusions

The combined tree suffers from the same problem as the smaller trees : the culled instances are only a small percentage of the whole dataset, so they are treated mainly as error, and the decision tree branches and rules for classifying culled instances are poor. In the whole dataset, only 14 culled instances have been classified correctly, leaving 949(!!) incorrectly classified as retained. However, the tree produced seems to make fairly logical branches, in terms of classifying the retained instances.

# 6. Classifying on PI Class

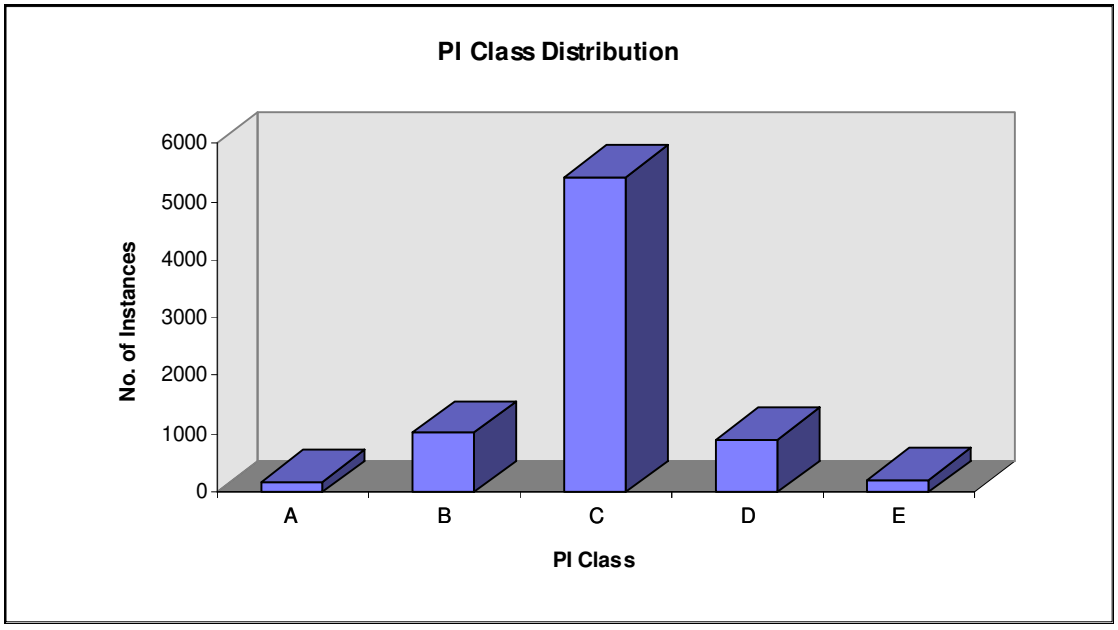
The production indices of an animal are a measure of it's actual production. The breeding index is a measure of the animal's genetic potential to produce. Therefore, the breeding indices of an animal should be good predictors of the animal's production indices. To test this theory, we devised a new attribute to classify the animals on.

Scatterplots of the production indices (Fat PI, Protein PI, Milk Volume PI and Payment PI) show that they have a high degree of linear correlation. To create the 'PI Class' attribute, a 'PI Score' was first calculated for each animal. This was taken as the mean of the animal's Fat PI, Protein PI and Payment PI. Then, the mean and standard deviation of the PI Score were found over all the animals. The PI Class was created based on the number of standard deviations the PI score was from the mean PI score.

## 6.1 PI Class Distribution

In the first run, the PI Class was based on the following divisions of PI Score;

| Std. Deviations from Mean PI Score | PI Class | Number of Instances |
|---|---|---|
| more than 2 | A | 171 |
| 1 to 2 | B | 1011 |
| -1 to 1 | C | 5426 |
| -2 to -1 | D | 900 |
| less than -2 | E | 212 |




The attributes included in this run were all derived from the BI values, as we were attempting to determine the effectiveness of the BI values to predict the PI's.

### 6.1.1 PI Class Tree

This tree was too large to display graphically

```
'WEKA_Payment_BI' <= 114.3 :
   'WEKA_Payment_BI' <= 110.6 : E (120.0/36.0)
   'WEKA_Payment_BI' > 110.6 : D (343.0/115.6)
'WEKA_Payment_BI' > 114.3 :
   'WEKA_Payment_BI' <= 133.7 :
      'WEKA_Payment_BI' <= 118.2 :
         'WEKA_Age' <= 8 :
            'WEKA_Payment_BI' <= 117.2 :
               'WEKA_MV_BI_Change' <= -0.1 : D (273.6/111.8)
               'WEKA_MV_BI_Change' > -0.1 :
                  'WEKA_Age' <= 6 : D (145.7/66.0)
                  'WEKA_Age' > 6 : C (80.7/37.4)
            'WEKA_Payment_BI' > 117.2 :
               'WEKA_AvgDiffPmt_BI' <= -2.994 : D (52.0/27.0)
               'WEKA_AvgDiffPmt_BI' > -2.994 : C (172.0/62.9)
         'WEKA_Age' > 8 :
            'WEKA_Protein_BI' <= 113.1 : D (58.0/31.1)
            'WEKA_Protein_BI' > 113.1 : C (149.0/25.7)
      'WEKA_Payment_BI' > 118.2 :
         'WEKA_Fat_BI' <= 140.5 :
            'WEKA_Payment_BI' <= 128.8 : C (4543.7/450.4)
            'WEKA_Payment_BI' > 128.8 :
               'WEKA_AvgDiffPmt_BI' > 13.369 : B (51.0/26.0)
               'WEKA_AvgDiffPmt_BI' <= 13.369 :
                  'WEKA_Age' <= 4 : C (492.0/125.2)
                  'WEKA_Age' > 4 :
                     'WEKA_MilkVolume_BI' <= 128.4 : C (147.0/65.7)
                     'WEKA_MilkVolume_BI' > 128.4 : B (52.0/19.9)
         'WEKA_Fat_BI' > 140.5 :
            'WEKA_AvgDiffFat_BI' > 12.221 : B (92.0/34.8)
            'WEKA_AvgDiffFat_BI' <= 12.221 :
               'WEKA_AvgDiffPrtn_BI' <= 1.862 : C (83.2/23.3)
               'WEKA_AvgDiffPrtn_BI' > 1.862 :
                  'WEKA_Protein_BI' > 128.6 : B (94.0/38.8)
                  'WEKA_Protein_BI' <= 128.6 :
                     'WEKA_Age' <= 3 : C (157.1/63.8)
                     'WEKA_Age' > 3 : B (122.0/49.3)
   'WEKA_Payment_BI' > 133.7 :
      'WEKA_Payment_BI' <= 138.5 : B (407.0/150.2)
      'WEKA_Payment_BI' > 138.5 : A (85.0/37.7)
```

| Before Pruning | | After Pruning | | |
|---|---|---|---|---|
| **Size** | **Errors** | **Size** | **Errors** | **Estimate** |
| 61 | 1477(19.1) | 41 | 1494(19.4%) | 20.7% |

| Classified as... | | | | | | |
|---|---|---|---|---|---|---|
| A | B | C | D | E | | |
| 51 | 111 | 9 | | | A | Actual Class |
| 34 | 524 | 453 | | | B | |
| | 183 | 5002 | 241 | | C | |
| | | 307 | 561 | 32 | D | |
| | | 22 | 102 | 88 | E | |

## 6.2 Conclusions

As expected, the BI values are reasonably accurate predictors of the PI Class - the tree shows that animals with higher BI values tend to have higher PI values (PI class) as well. While there is a high error rate in the tree (approximately 20%), the confusion matrix shows that almost all of these errors involve classifying an instance as one class too high or too low. The BI is only an estimator of the PI, so while a high correlation was expected, 100% correlation was not.

# 7. Overall Conclusions

## 7.1 Pre-processing of the Data

Early on, it became apparent that the original raw cow-culling datasets had characteristics which would produce misleading results from the Machine Learning Schemes being used. This was because the datasets were extracts of a relational database, whose fields had been constructed to be readable and understandable by LIC staff and custom software. The machine learning schemes required the data to have logical meaning which in some cases was quite different than the physical representation of the data provided, and a lot of effort was expended in transforming the data into this format.

Also present in the datasets were attributes which were dependent on other attributes, and non-linear attributes which combined to give a single meaning. We derived a number of new attributes from the original ones, while being careful not to obscure the meaning of the data. We were not what could be called "domain experts" who had professional understanding of the datasets, and this made our task somewhat more difficult.

## 7.2 Possible Inconsistency in the Data

As the cow-culling datasets are "real world" datasets, collected from real farms owned by real farmers, there is likely to be some inconsistency in the data. For example, some of the possible classifications for the 'Cause of Fate' attribute overlap. An animal labelled in one class by one farmer may be labelled in another by another farmer.

## 7.3 The Problem of Small Disjuncts

The cow-culling datasets are an example of a dataset in which one class comprises only a small percentage of the total number of instances. For instance, in the combined dataset of all participants, comprising a total of 16615 records, only 963 of these records are of cows that have been culled, while the remainder have been retained. This is partly because cows which are retained in multiple years generate multiple retained instances while those culled generate only one culled instance, and partly because only a small percentage of animals are culled each year anyway.

Small disjuncts are hard to classify using inductive learning schemes, which use an information gain heuristic based on the number of instances which can be correctly classified at each iteration. A small disjunct class will pose no problems if well defined classification patterns exist in the data, but if these patterns are not well-defined, then the class will have little influence on the information gain in the initial steps.

Although C4.5 is a robust learning scheme which can cope with a broad range of problems such as missing values and noise in the data, because the culled instances were only 5% of the total number of instances, C4.5 treated this entire class as noise in the data, as it fell within C4.5's parameters for acceptable error.

With the cow-culling datasets, we have found that while reasonably well-defined patterns exist in the retained instances, they do not appear to exist in the culled instances, at least within the given attributes. This may mean;

- Well defined patterns do not exist for culling cows (which seems unlikely)

- More attributes are needed in order to define the patterns

- The patterns exist in the data, but in a form unrecognisable by machine learning schemes - more attributes need to be derived in order to identify the patterns

## 7.4 Culling/Retaining Decisions

The trees induced from the datasets produce coherent rules for retaining cows with high production or breeding scores, but failed to produce coherent rules for culling cows which could successfully classify more than a small percentage of the culled instances. However, those rules which were induced were consistent with our expectations of what culling decisions might be based on. Other factors, possibly not present in the datasets, must also have an effect. Also, the datasets were based largely on production information, and therefore not apt to classifying animals which had not yet begun production.

A possible set of generic culling rules, constructed by hand from the output of the C4.5 runs;

If Bad Event then died
Otherwise, if age is 1 or less then retained
Otherwise, if Fat PI is greater than *Threshold1* then retained
Otherwise if Milk Volume PI is less than *Threshold2* then
    If Milk Volume PI is greater than *Threshold3* then retained
    Otherwise culled
Otherwise culled

*Thresholdn* is a value determined by the farmer's preferences and the characteristics of the herd.

# 8. References

***World Wide Web URL for The WEKA Machine Learning Project***

http://www.cs.waikato.ac.nz:80/~ml/

***References for this document***


[1] DeWar R.E., Neal D.L. (1994). *WEKA Machine Learning Project : Cow Culling*. Working Paper 94/12, University of Waikato, Hamilton, New Zealand.

[2] McQueen R.J., Neal D.L., DeWar R.E., Garner S.R. (1994). *Preparing and Processing Relational Data Through the WEKA Machine Learning Workbench*. University of Waikato, Hamilton, New Zealand.

[3] McQueen R.J., Neal D.L., DeWar R.E., Garner S.R., Nevill-Manning C.G. (1994). "The WEKA Machine Learing Workbench: Its Application to a Real World Agricultural Database". In *Proceedings of the Canadian Machine Learning Workshop*, Banff, Alberta, Canada.

[4] Monk T.J., Mitchell R.S., Smith L.A., Holmes G. (1994). "Geometric Comparison of Classifications and Rule Sets". In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases* pp. 395-406, Seattle, Washington.

[5] Quinlan J.R. (1986). "Induction of Decision Trees". *Machine Learning 1 (1)*, pp 81-106.

[6] Quinlan J.R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California.

[7] Witten I.H., Cunningham S.J., Holmes G., McQueen R.J., Smith L.A. (1993). "Practical Machine Learning and its Potential Application to Problems in Agriculture". In *Proceedings of The New Zealand Computer Conference* vol. 1 pp. 308-325, Auckland, New Zealand.