Implementation of Data Mining in Estimating The Growth Of Local Sheep

Aan Kardiana[1], Lilis Khotijah[2]

[1] Faculty of Information Technology, YARSI University
Jakarta, 10510, Indonesia

[2] Faculty of Animal Science, Bogor Agricultural University
Bogor, 16680, Indonesia

Abstract

Data mining is a process to use statistical technique, mathematics, artificial intelligence, and learning machine to extract, identify beneficial information and discovery knowledge from database. In this research, the authors apply this method to estimate the growth of local sheep. Research method consists of several phases, namely: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evolution and Knowledge Presentation. Data as amount of 4357 samples, processed by using CART (Classification and Regression Tree) and Correlation Analysis method. The Average Daily Gain is target variable is and indicator variable consist of dry matter intake from : Grass; Corn; Cassava Meal; Coconut Meal; $CaCO_3$; Salt; Premix; Urea; Corn Oil; Corncob; Soybean Meal; Fish Meal and Sunflower Oil. The knowledge presentation gotten is Coconut Meal as dominant indicator variable. The optimal regression trees that has 41 terminal nodes with relative error of 0,659, can be used to determine composition ingredient base on daily gain expected.

Keywords: Data mining, regression tree, estimation, average daily gain

1. Introduction

Beef Self Sufficient Program 2010 is a government programs to supply animal protein in order to feed security. Until now, beef production ability is just able to give contribution around 70-75% from national needs, whereas government launch beef production role can give contribution around 90-95% from national needs. Mutton and lamb in Indonesia only reaches 0.24 g. It is still very low than in several other countries, such as German of 3.33 g, Russia of 3.36 g, and China of 6.36 g. Those numbers will increase continuously in line with the increase of population and awareness level upon the importance of animal protein for nation intelligence [1].

Efforts to increase sheep role as contributor of qualified animal protein source is significantly determined by its productivity level. A lot of researches in livestock field have been done to discover sheep potency and increase that productivity. Therefore, there are many research data collected, but those are not yet utilized optimally now. Current processing method is tabulation and parametric statistic (regression, correlation, and variance analysis). Publication of processing result is still limited only in environmental science farm, whereas many data generated from many researches can be information source not only for livestock field, but can be useful for other related knowledges either directly or indirectly.

Several problems faced are: data generated from researches is quite big so it needs big database; Research results in this field are still partially connected, not yet comprehensively integrated to use for developing livestock sector.

To respond above mentioned problems, new processing methods that can process big data and integrate research results are needed. Another approach that can be used is data mining. This research will apply Data Mining method in collected research data to find valuable hidden information which can be used in developing livestock sector.

2. Literature Study

2.1 Data mining

Turban *et al.* [2] defines data mining as process to use statistical technique, mathematics, artificial intelligence, and learning machine to extract and identify related beneficial information and knowledge from any big database.

Data mining is an essential step in the process of knowledge discovery, consists of an iterative sequence of the following steps [3]:

1. Data cleaning, to remove noise and inconsistent data.
2. Data integration, where multiple data sources may be combined.

3. Data selection, where data relevant to the analysis task are retrieved from the database.
4. Data transformation, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining, an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation, to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Larose [4] expresses that the task of data mining are:
1. Description. Simply researcher want to find ways to describe pattern and trend existing in data.
2. Estimation. Estimation model is developed by using complete data that contains value from target variable as prediction value. Then, based on value substitution of prediction variable, it is known that estimation model resulted can known target variabel value. Target variable as numerical.
3. Classification. Classification has categorical target variable.
4. Prediction. Prediction is almost the same with estimation and classification, except in prediction, value from result variable will be exist in the future.
5. Cluster. Constitute data group that has similarity.
6. Association. Find attribute that appear simultaneously.

Data mining task that will be done in this research are description and prediction by using regression tree method.

Breiman *et al.* [5] expresses that regression tree is partitioned by a sequence of binary splits into terminal nodes. In each terminal node t, the predicted response value y(t) is constant. Regression tree formation phases are:
1. Growing the initial tree
   The initial tree is grown through phase:
   a. Select root node.
   b. Determine all splits that might be formed from all indicator variable and calculate homogeneity level.
   c. Select the best indicator variable that has the highest homogeneity level.
   d. Do changing on other branch node.
   e. Stop growing the tree if there is no change on homogeneity level significantly.
2. Determine optimal tree
   The initial tree that has been formed has big size, as a result of using tree formation stop criteria. It is difficult to present the knowledge. To avoid estimation of overfitting, the prunning process is done use the 10-cross validation sample therefore optimal tree is generated.

2.2 Livestock Productivity

Livestock productivity is determined by consumption value of food substance, the increase of body weight and effectiveness to use feed. The increase of weight constitutes ability from animal to change food substances contained in feed to form muscle tissue (meat) that can be known by repeated weighing every day, week or month [6]. Food consumption value is total food consumed by animal if they are given adlib. efficiency to give feed constitutes ration between total feed consumed with total increase of body weight generated [7].

3. Result and Discussion

Data collected is 4357 data [8] used to develop regression tree that use CART (Classification and Regression Tree) method supported by Salford Predictive Modelling (SPM) software issued by Salford System [9].

The Average Daily Gain is target variable, while indicator variables consist of dry matter intake from: Grass; Corn; Cassava Meal; Coconut Meal; CaCO3; Salt; Premix; Urea; Corn Oil; Corncob; Soybean Meal; Fish Meal and Sunflower Oil.

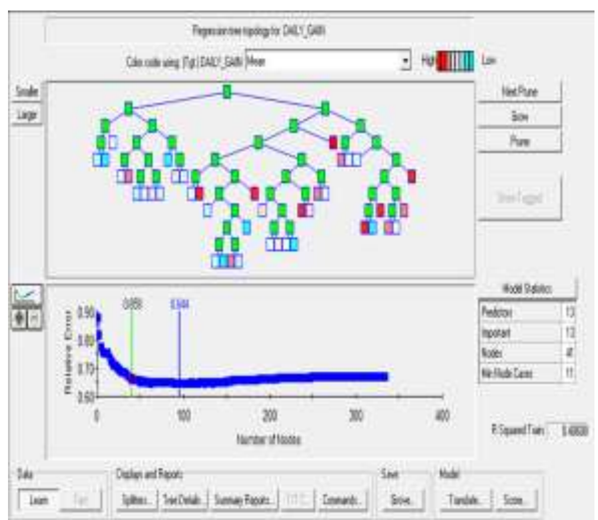
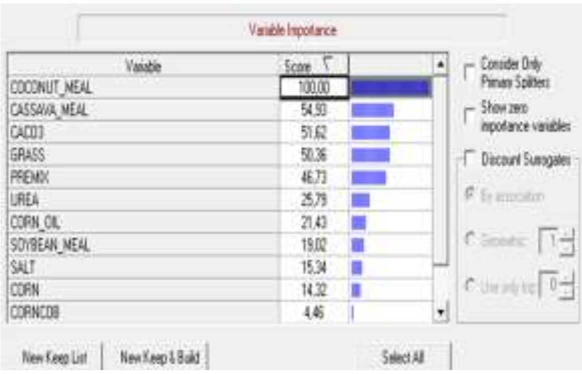


Fig. 1 Regression Tree Toplogi

Figure 1 shows that optimal regression tree has 41 terminal nodes with relative error is 0,659 and involves 10 indicator variables.

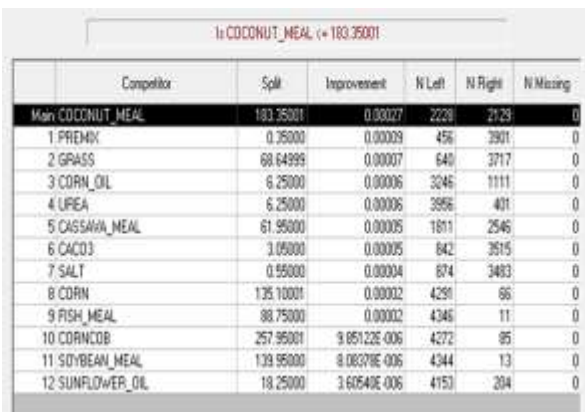
The dominant indicator variable is dry matter intake from Coconut Meal. This variable becomes the best split on root node, with the highest Variable Importance and Improvement value among 12 other variables (Figure 2 and Figure 3).



Variable Importance

Variable	Score
COCONUT_MEAL	100.00
CASSAVA_MEAL	54.93
CACO3	51.62
GRASS	50.36
PREMIX	46.73
UREA	25.79
CORN_OIL	21.43
SOYBEAN_MEAL	19.02
SALT	15.34
CORN	14.32
CORNCOB	4.46

Fig. 2 The Variable Importance



1: COCONUT_MEAL <= 183.35001

	Competitor	Split	Improvement	N Left	N Right	N Missing
Main	COCONUT_MEAL	183.35001	0.00027	2228	2129	0
1	PREMIX	0.35000	0.00009	456	3901	0
2	GRASS	68.64999	0.00007	640	3717	0
3	CORN_OIL	6.25000	0.00006	3246	1111	0
4	UREA	6.25000	0.00006	3956	401	0
5	CASSAVA_MEAL	61.95000	0.00005	1811	2546	0
6	CACO3	3.05000	0.00005	842	3515	0
7	SALT	0.55000	0.00004	874	3483	0
8	CORN	135.10001	0.00002	4291	66	0
9	FISH_MEAL	88.75000	0.00002	4346	11	0
10	CORNCOB	257.95001	9.85122E-006	4272	85	0
11	SOYBEAN_MEAL	139.95000	8.08378E-006	4344	13	0
12	SUNFLOWER_OIL	18.25000	3.60540E-006	4153	204	0

Fig. 3 Root Splits

This result is in line with Pearson Correlation Coefficient between indicator variable with target variable as stated in Table 1. Coconut Meal variable has the highest Pearson Correlation Coefficient value with Average Daily Gain variables (0,405) among other indicator variables.

This is in line also with analysis result that indicates Coconut Meal has the highest protein content among other food material sources, where protein constitutes main food to form tissue in the infancy. This is also in accordance with NRC [10] that one of factors influences average daily gain is total protein consumed everyday.

Root node is splitted by dry matter intake from Coconut Meal variable, if less than or the same with 183,35001 g split to node 2 and if more than 183,35001 g is split to node 3 (Figure 4).

Table 1: Correlation Coeffisien with Indicator Variable and Average Daily Gain

Correlation Coeffisien	***Average Daily Gain***
Coconut Meal	0,405
Grass	0,152
Corn Oil	0,144
Cassava Meal	0,059
Salt	0,054
Fish Meal	0,024
Soybean Meal	0,020
Corn	0,015
Premix	0,014
Sunflower Oil	0,011
Corncob	0,006
CaCO3	0,005
Urea	-0,073

Node 2 and node 3 are developed become next nodes based on splits that has the highest Variable Importance and Improvement value on those nodes.

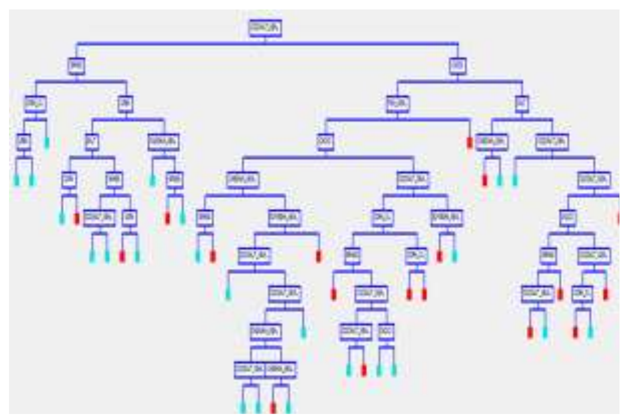


Fig. 4 Optimal Tree

The biggest tree has 335 terminal nodes and after pruning by using 10-Cross Validation, optimal tree is gotten with 41 terminal nodes. Figure 4 indicates whatever indicator variable contained in optimal regression tree that also involved in estimation of body weight increase value as target variable. This tree gives knowledge representation concerning combination of whatever indicator variables that can be used to determine composition ingredient base on daily gain expected.

Recommendation concerning estimation of food material content that can be used in accordance with the increase of average daily gain expected based on rule on terminal nodes are if dry matter intake value from ingredient:

```
/*Terminal Node 14*/
if
(
   COCONUT_MEAL > 183.35 &&
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   CASSAVA_MEAL <= 104.5 &&
   GRASS > 151.05
)

{
   terminalNode = -14;
   mean = 0.152231
}

/*Terminal Node 18*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   SOYBEAN_MEAL <= 139.95 &&
   COCONUT_MEAL > 227.8 &&
   COCONUT_MEAL <= 400.7 &&
   CASSAVA_MEAL > 172 &&
   CASSAVA_MEAL <= 185.05
)

{
   terminalNode = -18;
   mean = 0.122449
}

/*Terminal Node 21*/
if
(
   COCONUT_MEAL > 183.35 &&
   FISH_MEAL <= 88.75 &&
   CACO3 <= 3.15 &&
   CASSAVA_MEAL > 104.5 &&
   SOYBEAN_MEAL > 139.95
)

{
   terminalNode = -21;
   mean = 0.138462
}

/*Terminal Node 27*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 > 3.15 &&
   CACO3 <= 6.35 &&
   COCONUT_MEAL > 183.35 &&
   COCONUT_MEAL <= 258.45 &&
   CORN_OIL > 6.55 &&
   CORN_OIL <= 8.3
)

{
   terminalNode = -27;
   mean = 0.1449
}

/*Terminal Node 29*/
if
(
   FISH_MEAL <= 88.75 &&
   CACO3 > 3.15 &&
   CACO3 <= 6.35 &&
   COCONUT_MEAL > 258.45 &&
   SOYBEAN_MEAL <= 29.6
)

{
   terminalNode = -29;
   mean = 0.12377
}

/*Terminal Node 31*/
if
(
   COCONUT_MEAL > 183.35 &&
   CACO3 <= 6.35 &&
   FISH_MEAL > 88.75
)

{
   terminalNode = -31;
   mean = 0.163633
}

/*Terminal Node 35*/
if
(
   SALT > 0.95 &&
   CACO3 > 6.35 &&
   CACO3 <= 6.85 &&
   GRASS <= 133.45 &&
   COCONUT_MEAL > 207.4 &&
   COCONUT_MEAL <= 220.5
)

{
   terminalNode = -35;
   mean = 0.149837
}
```

```
/*Terminal Node 37*/
if
(
   SALT > 0.95 &&
   COCONUT_MEAL > 207.4 &&
   COCONUT_MEAL <= 387.45 &&
   CACO3 > 6.35 &&
   CACO3 <= 6.85 &&
   GRASS > 133.45
)

{
   terminalNode = -37;
   mean = 0.153565
}

/*Terminal Node 40*/
if
(
   SALT > 0.95 &&
   CACO3 > 6.85 &&
   COCONUT_MEAL > 314.6 &&
   COCONUT_MEAL <= 387.45
)

{
   terminalNode = -40;
   mean = 0.130885
}
/*Terminal Node 41*/
if
(
   CACO3 > 6.35 &&
   SALT > 0.95 &&
   COCONUT_MEAL > 387.45
)

{
   terminalNode = -41;
   mean = 0.149706
}
```

That means, if CACO3 > 6.35 g and SALT > 0.95 g and COCONUT_MEAL > 387.45g, so AVERAGE DAILY GAIN will be 149.706 g.

4. Conclusions

From this research we conclude that :

1. Implementation of data mining in estimating the growth of local sheep generates maximum size regression tree that contains 335 terminal nodes.
2. The dominant indicator variable is dry matter intake from Coconut Meal.
3. The optimal regression tree that has 41 terminal nodes with relative error of 0,659 can be used to determine composition ingredient base on average daily gain expected.

Acknowledgments

This study is one of research roadmap of the Faculty of Information Technology YARSI University, and funded by the Directorate General of Higher DIPA Education Ministry of National Education through Grant named "Hibah Unggulan Perguruan Tinggi".

References

[1] Heriyadi D. Domba dan kambing di Indonesia: Potensi, Masalah dan Solusi, Majalah TROBOS No. 101 Februari 2008 Tahun VIII, 2008.

[2] Turban E, Aronson JE, Liang TP. Decision Support Systems and Intelligent Systems, Seventh 7/E, Prentice Hall, 2005.

[3] Han J, Kamber M. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publiser, Elsevier, San Francisco, 2006.

[4] Larose DT. Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons Inc, New Jersey, 2005.

[5] Breiman L, Friedman JF, Olshen RA, Stone CJ. Classification and Regression Tree, Chapman & Hall Inc, New York, 1993.

[6] Tillman AD, Reksohadiprodjo S, Prawirokusumo S, Hartadi H, Lebdosoekojo S. Ilmu Makanan Ternak Dasar, Gadjah Mada University Press, Yogyakarta, 1998.

[7] Parakkasi A. Ilmu Nutrisi Ternak Ruminansia, UI Press, Jakarta, 1995.

[8] INTDK. Kumpulan Data Hasil Penelitian Nutrisi Ternak Domba 2010-2012, Laboratorium Ilmu Nutrisi Ternak Daging dan Kerja, Departemen Ilmu Nutrisi dan Teknologi Pakan, Fakultas Peternakan, IPB, 2012.

[9] Steinberg D, Golovnya M. CART® 6.0 User's Manual, San Diego, CA: Salford Systems, 2006.

[10] NRC. Nutrient Requirements of Small Ruminant. The National Academies Press, Washington Dc, 2006.

Aan Kardiana holds BSc from Bandung Institute of Technology and MSc from IPB (Bogor, Indonesia) in 2000. He is currently an academic, research staff and Head of Computational Intelligent Research Group of Faculty of Information Technology, YARSI University. His research interests are data mining, statistics and e-Health. He is also a member of YARSI E-Health Research Center (YEHRC); has won some research grants; published a number of papers in national proceeding and international journal.

Lilis Khotijah holds BSc and MSc from Faculty of Animal Science from from Bogor Agricultural University (IPB) in 1999. She is currently an academic, lecturer of Animal Nutrition and research staff of Faculty of Animal Science, Bogor Agricultural University. Her research interest is Nutrition Reproduction of Animal.