Abstract #:51557

# Prediction of Pregnancy Outcome Using Machine Learning Algorithms.

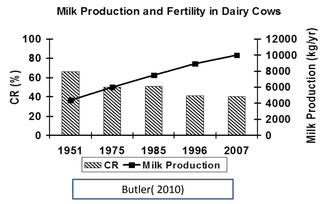
Saleh Shahinfar[1], Kent Weigel[1], David Page[2], Jerry Guenther[1], Victor Cabrera[1], Paul Fricke[1].

[1]Department of Dairy Science, University of Wisconsin-Madison

[2]Department of Biostatistics and Medical Informatics and Department of Computer Science, University of Wisconsin-Madison

shahinfar@wisc.edu

## Here is the Problem :




| Milk Range, kg/cow | Number of Herds | Avg Milk/cow, kgs. | CR %, 1st to 3rd service |
|---|---|---|---|
| 8165 - 9071 | 262 | 8661 | 45 |
| 9072 - 9978 | 299 | 9536 | 42 |
| 9979 - 10886 | 259 | 10396 | 38 |
| 10887 - 11793 | 165 | 11276 | 36 |
| 11794 - 12700 | 64 | 12176 | 36 |
| > 12701 | 25 | 13365 | 32 |

**During the last few decades, successful selection for milk production in dairy cattle has resulted in a dramatic decrease in the reproductive ability of dairy cows.**

## Introduction and Objective

On a daily basis, making a more informed decision about whether or not to breed a given cow based on knowledge about the expected outcome of the breeding would enhance profitability of breeding program and increase net income of the farm. The outcome of each breeding can be affected by many management and physiological factors that vary widely between farms. Machine learning algorithms offer a great opportunity with regard to problems of multicollinearity, missing values, or complex interactions among variables (Caraviello et al., 2006). The objective of this study was to develop a user friendly and intuitive on-farm tool to help farmers make decisions about breeding specific cows. In order to achieve our goal, we applied several different machine learning algorithms to predict the pregnancy status of each cow after breeding based on phenotypic and genotypic data.

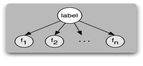
The objective of this study was to compare the performance of different machine learning algorithms based on their ability to predict pregnancy status in dairy cattle using a large data set comprised of production , reproduction, health, and genetic information. Subsequently, we used the best-performing algorithm to do a cost/benefit evaluation and show how the optimal decision changes in different scenarios based on pregnancy values, breeding costs, and days open costs.

## Material and Methods

Data: Included 26 dairy farms in the Alta Genetics Advantage Progeny Testing Program from 2000 to 2010 with reproduction, production, health events, and breeding values of cows and sires. A total of 129,245 breeding records and 28 explanatory variables were available for primiparous cows, and 195,128 breeding records and 31 explanatory variables were available for multiparous cows.

Methods:

**Naïve Bayes classifier** is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Even though the independency assumption between features is not true in reality, it will outperform many other elaborate, complex classifiers.
Naïve Bayes is the simplest form of Bayesian network in which all attributes are independent given the values of the class variables (Zhang, 2004).

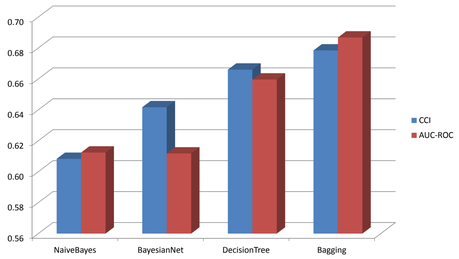



**Bayesian network** represents the joint probability distribution of set of variables{ $X_1$, $X_2$, …. $X_n$} as a discrete acyclic graph and a set of conditional probability distributions corresponding to specific features.

**Decision Trees** are one of the simplest, most intuitive, easily interpretable, and widely used machine learning algorithms. They used information theory as measurement for divide and conquer approach to come up with a classifier. Each node in Decision Tree is a variable which divide instances that reach to that node by a condition. Leaves in Decision Tree are assignment of class labels.

**Bagging** stands for bootstrap aggregation and is one of the ensemble methods. It is a method for generating multiple versions of a predictor and using these to get aggregated predictors. For predicting numeric values, aggregating would be a simple average over all models, and for classification it would be the majority of votes (Breiman, 1996).

## Results



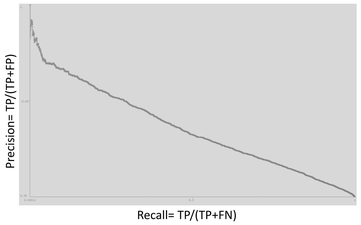
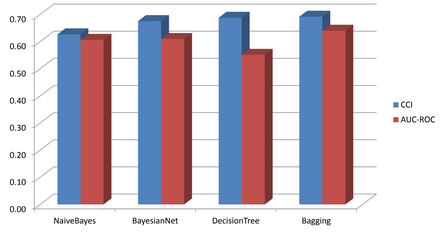




**Figure 1: Comparison between performance (%) of four different machine learning algorithms for predicting pregnancy outcome in primiparous cows, (CCI% = correctly classified instances, AUC-ROC% = area under curve - receiver operating characteristic).**

**Table 1: t-test evaluation of the difference in predictive ability between bagging and alternative methods in primiparous cows.**

| | Bagging vs. Naïve Bayes | | Bagging vs. Bayesian Network | | Bagging vs. Decision Tree | |
|---|---|---|---|---|---|---|
| Error Criteria | P-value | 95% CI | P-value | 95% CI | P-value | 95% CI |
| CCI | <0.001 | 0.06 , 0.08 | <0.001 | 0.03 , 0.05 | 0.04 | 0.00, 0.02 |
| AUC-ROC | <0.001 | 0.06 , 0.09 | <0.001 | 0.06 , 0.09 | <0.001 | 0.01, 0.04 |



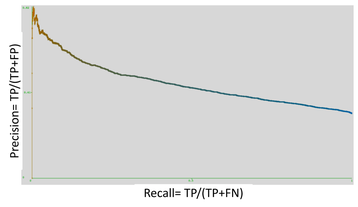




**Figure 2: Comparison between performance (%) of four different machine learning algorithms for predicting pregnancy outcome in multiparous cows, (CCI = correctly classified instances, AUC-ROC = area under curve - receiver operating characteristic).**

**Table 2: t-test evaluation of the difference in predictive ability between bagging and alternative methods in multiparous cows.**

| | Bagging vs. Naïve Bayes | | Bagging vs. Bayesian Network | | Bagging vs. Decision Tree | |
|---|---|---|---|---|---|---|
| Error Criteria | P-value | 95% CI | P-value | 95% CI | P-value | 95% CI |
| CCI | <0.001 | 0.06 , 0.07 | <0.001 | 0.01 , 0.02 | 0.32 | -0.001, 0.01 |
| AUC-ROC | <0.001 | 0.02 , 0.04 | <0.001 | 0.02 , 0.04 | <0.001 | 0.06, 0.12 |

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2010-85122-20612 from the USDA National Institute of Food and Agriculture.

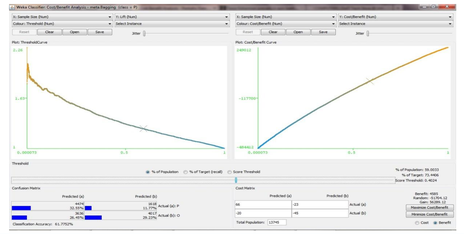
## Cost/benefit analysis

Table 3: Pregnancy value for primiparous cows by days in milk and milk yield class (Kalantari et al, 2010).

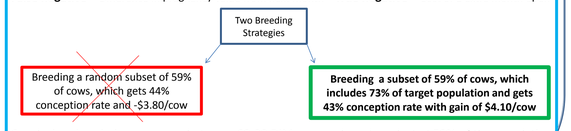
| | Milk Yield Class (MC) | | |
|---|---|---|---|
| Days in Milk | MC2 | MC3 | MC4 |
| 60-90 | $ 66 | $ 65 | $ 63 |
| 90-120 | $ 89 | $ 87 | $ 86 |
| 120-150 | $ 124 | $ 123 | $ 122 |
| 150-180 | $ 149 | $ 148 | $ 147 |

Table 4: Days open cost for primiparous cows by days in milk and milk yield class (Kalantari et al, 2010).

| | Milk Yield Class (MC) | | |
|---|---|---|---|
| Days in Milk | MC2 | MC3 | MC4 |
| 60-90 | $ -45 | $ -43 | $ -40 |
| 90-120 | $ -54 | $ -52 | $ -52 |
| 120-150 | $ -71 | $ -72 | $ -72 |
| 150-180 | $ -78 | $ -79 | $ -80 |



Assumptions for costs/benefits classification:
**True positive** = Pregnancy value
**False positive** = Breeding cost ($20)
**False negative** = Difference in pregnancy value from next month
**True negative** = Cost of 1 extra month open




Conclusion: In primiparous cows between 60-90 DIM, we can breed a selected 59% of the population that includes 73% of the target (TP) group, and we will gain $56,289 in comparison with breeding a randomly chosen 59% of the population.

## Conclusion

1- An Information-based variable selection procedure identified mean conception rate in last 3 month in the herd, period, DIM at breeding, past times bred, current times bred, and calving interval as the most effective explanatory variables in predicting pregnancy outcome.

2- Bagging (with decision trees) had significantly higher performance compared with the other three methods in both primiparous and multiparous cows. For multiparous cows, the performance of bagging and decision tress did not significantly in CCI, but bagging was significantly better according to AUC-ROC.

3- Cost/benefit evaluations should be carried out on the classification results in order to determine the financial gain that farmers can achieve by selecting the most optimal subset of cows for breeding.

## References

Kalantari, A. S., H. Mehrabani-Yeganeh, M.Moradi, A. H. Sanders, and A. De Vries. 2010. Determining the optimum replacement policy for Holstein dairy herds in Iran. J. Dairy Sci. 93:2262-2270.
Caraviello D.Z, K.A. Weigel, M.Craven, D.Gianola, N.B. Cool, K.V. Nordlund, P.M. Fricke and M.C. Wiltbank. 2006. Analsis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. J Dairy Sci. 89:4703-4722.
Breiman, L. 1996. Bagging predictors. Machine Learning, 24:123-140.
Zhang, H., 2004. The optimality of Naïve Bayes. In Proc. 17th Internat. FLAIRS Conf., Florida, USA.
Butler, W.R. 2010. Energy balance in the transition cow and reproduction. Proceedings Dairy Cattle Reproduction Council (DCRC) Annual Meeting, St. Paul, MN, pp. 93-100.