ÁRVORE DE DECISÃO APLICADA EM DADOS DE INCUBAÇÃO DE MATRIZES DE POSTURA HY-LINE W36

Decision tree applied to hatchery databases of Hy-Line W-36

Marcelo Gomes Ferreira Lima[1], Luiz Henrique Antunes Rodrigues[2]

RESUMO

Incubatório de ovos é um setor de grande importância na Avicultura de postura. Com a redução dos custos dos equipamentos de informática cresce o armazenamento de dados para gerenciamento do processo produtivo. A Mineração de Dados surge como uma técnica para identificar conhecimentos novos e úteis nos bancos de dados. Objetivou-se, neste trabalho, explorar a técnica Arvore de Decisão em banco de dados de incubatórios de matrizes de postura, visando a elaboração de padrões de incubação. Foram disponibilizados, pela empresa Hy-Line do Brasil Ltda, dados de incubação entre os anos de 2002 e 2006 da linhagem Hy-Line W-36. Dois experimentos foram realizados. Em um deles, valores acima dos estabelecidos pela empresa como desejado para o índice "fêmeas nascidas vendáveis" foram identificados como relevantes para a geração das regras. No outro, valores abaixo dos estabelecidos pela empresa foram identificados como relevantes para a geração das regras. Foi utilizado o algoritmo Entropia C 4.5 e o *software* SAS-Enterprise Miner como ferramenta de análise . Como conclusão deste estudo, foi possível observar que com a técnica estudada, os dados utilizados no gerenciamento de produção são suficientes para identificar conhecimentos novos, úteis e aplicáveis a fim de melhorar a produtividade das empresas incubadoras, atendendo à demanda com diminuição do desperdício.

Termos para indexação: Mineração de dados, KDD, inteligência artificial, avicultura.

ABSTRACT

Hatchery is a very important sector in egg production. As computers become cheaper, there is an increase in data storage for the production management process. Data Mining has appeared as a technique to identify new and useful knowledge in databases. The objective of this work was to explore the Decision Tree technique in hatchery databases to identify the best standards of the incubation process. The data set used in this research was supplied by Hy-Line do Brasil Ltda., corresponding to the incubation period of 2002-2006, from the strain Hy-line W-36. Two experiments were carried out. In the first experiment, values higher than the company's standards for saleable females were identified as relevant to generate the rules. In the second experiment, values below those established by the company were identified as relevant for the generation of rules. Entropy C 4.5 algorithm and the software SAS-Enterprise Miner were used for data analysis. The conclusion is that, with the technique studied, the data used for production management are sufficient to identify new, useful and applicable knowledge in order to increase productivity of hatcheries, catering for the demand with less waste.

Index terms: Data mining, KDD, artificial intelligence, poultry science.

(Recebido em 15 de janeiro de 2009 e aprovado em 20 de abril de 2010)

INTRODUÇÃO

A necessidade por melhores resultados de produção faz o setor avícola buscar inovações constantemente. A demanda por sistemas de gerenciamento de informações que ofereçam recursos para tomada de decisão vem crescendo à medida que a complexidade da avicultura mundial aumenta. Incubatórios avícolas são as empresas responsáveis por reproduzir o material genético. Como outros setores de produção, os incubatórios geram diariamente dados que são utilizados para o gerenciamento das atividades.

Diversas Técnicas de Mineração de Dados vêm sendo utilizadas para detectar relacionamento entre diferentes atributos em grandes bancos de dados (Fayyad et al., 1996). Utilizando a técnica Redes Neurais Artificiais, Roush et al. (1996), obtiveram sucesso na identificação de padrões de ocorrência de ascite em frangos. Piettersma et al. (2003) estudaram o uso da técnica Árvore de Decisão para analisar curvas de lactação de vacas de leite, facilitando a seleção de animais mais produtivos. Da mesma forma, (Kirchner et al., 2004), aplicaram essa técnica para identificar padrões de produção em suínos para seleção de fêmeas reprodutoras, permitindo a antecipação da seleção das mesmas.Também utilizando a técnica de Redes Neurais Artificiais, Oliveira et.al.(2010), validou seu uso da para fazer a previsão da produção de álcool utilizando dados da série hostórica.

O resultado de incubação pode ser afetado por vários fatores, entre eles o tamanho do ovo, influenciado

[1]Agroinfo Tecnologia da Informação Ltda – Campinas, SP
[2]Faculdade de Engenharia Agrícola – 13083-875 – Campinas, SP – lique@feagri.unicamp.br

pela idade da matriz (Vieira & Moran, 1998), e tempo de estocagem do ovo (Elibol et al., 2002).

Objetivou-se, nesta pesquisa, padrões de incubação dos lotes de aves da linhagem Hy-Line W36, que tiveram os melhores e os menos favoráveis resultados de fêmeas vendáveis utilizando a técnica Árvore de Decisão.

MATERIAL E MÉTODO

O presente estudo foi conduzido utilizando dados de gerenciamento de incubação de incubatório comercial, no período de agosto de 2002 a setembro de 2006. Estavam disponíveis 10.300 registros de dados de incubação. As aves estavam alojadas em galpões com ambiente automaticamente controlado visando à minimização dos efeitos da temperatura externa, priorizando o conforto térmico dos animais. Os dados de gerenciamento de incubação utilizados foram: data de incubação, linhagem da poedeira, idade da poedeira no dia da postura, idade do ovo no dia da incubação, total de ovos incubados, % de fêmeas vendáveis (pintinhas em condições de venda / total de aves nascidas do lote) e % de fêmeas vendáveis padrão (valores tabulados pela empresa - *Hy-line)*. Buscando-se identificar possível influência da temperatura externa no dia da postura sobre os resultados de incubação, foram adicionados dados climáticos, temperatura ambiente mínima e máxima do dia da postura, Também foi incluida a incidência de luz, tendo por base estudos de FREITAS et al. (2005). A partir dos atributos disponibilizados foram criados novos atributos : Número de Lotes Incubados no Dia, Idade Média dos Ovos na Incubação, Total de Ovos incubados no Dia, Amplitute térmica de Temperatura (temperatura máxima – temperatura mínima) definida como AT, Conforto Térmico (identificação de dia com temperatura máxima acima do conforto térmico) e Dias Consecutivos em Desconforto Térmico.

A Árvore de Decisão (AD), técnica de Mineração de Dados escolhida para identificar padrões de produção, por ser uma modelagem preditiva, exige a definição de um atributo Alvo ou preditor. Esse atributo orienta o algoritmo AD a classificar os demais atributos, quando possível, para identificar padrões que se repitam no repositório de dados fornecido. Para tanto foi desenvolvido o atributo identificado como Delta cujo conteúdo é a diferença entre a % de fêmeas vendáveis produzidas e a % de fêmeas vendáveis tabulada (ou esperada). O valor calculado identifica os resultados abaixo ou acima do desejado.

Foram filtrados registros cujo valor de Delta estavam entre –10 e +10 pontos porcentuais de diferença, o que reduziu a base de dados a 6.361 registros.

Dois experimentos foram propostos. O primeiro, chamado POS, procurou identificar as características dos lotes de aves que tiveram Delta de incubação superior ao esperado. O segundo experimento, NEG, identificou as incubações que obtiveram os resultados abaixo do desejado, como apresentado na Figura 1.

A partir do atributo Delta apresentado, foi gerado um novo atributo Delta_STX, de conteúdo binário, assumindo valores S ou N, para identificar o resultado desejado de incubação em cada experimento.Conforme apresentado na tabela 1. Assim, no experimento POS, S são os registros com valores superiores e no experimento NEG são os registros com valores inferiores.

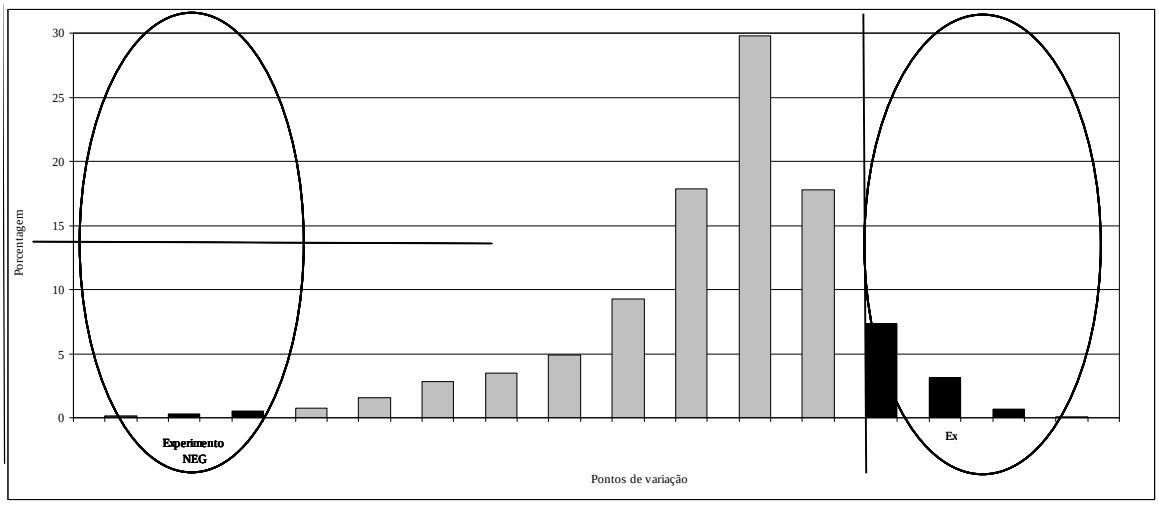

Figura 1 – Seleção dos registros de dados para os experimentos propostos.

A técnica AD foi gerada com o aplicativo SAS Enterprise Miner (SAS-EM), versão 4.3 para PC. Foi utilizado o módulo filtragem de dados (Tabela 2), para selecionar os registros da linhagem desejada. A técnica sugere o particionamento dos dados para análise em, no mínimo, dois grupos o de treinamento e o de validação. Foram separados 70% e 30% dos dados, respectivamente, para treinamento e validação, sendo utilizado para tanto o módulo *Data Partition*. Para utilização do módulo *Tree* (algoritmo AD), foram feitas as seguintes configurações: Algoritmo Entropia C4.5, *Split* binário, 10 registros mínimos por folha e 10 níveis de profundidade da AD.

Tabela 1 – Respectivos valores da classe **S** do atributo DELTA_STX.

Número de registros total	Experimentos			
	POS	% registros	NEG	% registros
6.361	>4,4	18,0	<= 2,0	21,0

Tabela 2 – Limites estabelecidos para os atributos filtrados.

Atributo	Importance
DELTA	-10,0 a +10,0
Temperatura Máxima	18^{o}C até o máximo encontrado nos registros
Idade dos Ovos	Entre 0 e 20 dias de idade
AT	Entre 18 e 42^{o}C

RESULTADOS E DISCUSSÃO

Como fase de pré-processamento dos dados, o software SAS-EM executa o módulo Seleção de Variáveis para identificar as variáveis de maior importância para a modelagem. O índice *Importance* é calculado para classificar os atributos. Para atributos do tipo numérico é calculada soma dos R^2 e para atributos categóricos a soma do índice Gini (SAS Institute, 2005). O atributo de valor máximo de *Importance* recebe o valor 1, e os demais atributos são comparados a ele. Atributos com *Importance* menor ou igual a 0,05, são automaticamente rejeitados e não são disponibilizados para execução do algoritmo.

Com a execução do algoritmo C4.5 são geradas regras (*English Rules)* ou padrão dos dados. As regras foram apresentadas ao especialista de incubação para validar o novo conhecimento. Para cada regra identificada foi calculada a respectiva acurácia (grau de acerto) da regra. O especialista interpretou que regras com acurácia igual ou superior a 50% são regras de potencial interesse.

Experimento POS

Na Tabela 3, são apresentados os resultados da seleção de variáveis do experimento POS.

Foram rejeitados os atributos Ovos incubados/lote, Temperatura ambiente acima do conforto térmico, Temperatura mínima e AT, por apresentarem baixo valor de *Importance*, ou seja, baixa correlação com o atributo preditor.

A Árvore de Decisão gerada pelo software SAS-EM para o experimento POS, linhagem Hy-Line W-36 está ilustrada na Figura 2.

Foram geradas sete regras de interesse com acurácia entre 54 e 86%. Na Tabela 4 são apresentadas as regras, no formato descritivo, facilitando o entendimento.

Tabela 3 – Atributos selecionados no módulo Seleção de Variáveis do SAS-EM para o experimento POS.

Atributos	**Linhagem W-36**	
	Importance	Ordem
Idade do lote - Idade lote	1,0000	1
Número de lotes incubados – Nr Lote Inc	0,2725	2
Idade Média Incubação - Id MD Inc (dias)	0,2644	3
Total de ovos incubados no dia – Total dia	0,2642	4
Idade ovo – IO	0,1999	5
Luz	0,1393	6
Dias com temperatura acima do conforto térmico-DACT	0,1088	7
Temperatura Máxima (^{o}C) – Max	0,1087	8
Número de ovos incubados do lote – Ovo Inc		

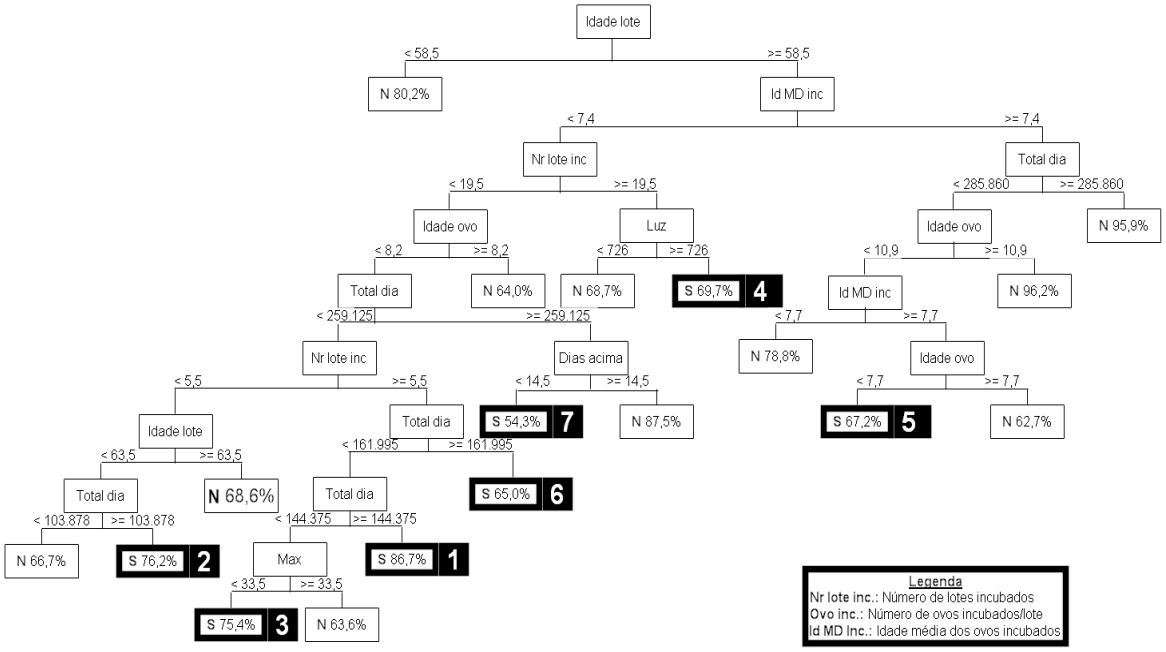

Figura 2 – Árvore de Decisão adaptada, experimento POS.

Tabela 4 – Regras de interesse geradas pelo SAS-EM para o Experimento POS para a linhagem Hy-Line W-36.

Nro. Reg.	Acurácia (%)	IO	Nr Lote Inc	Max	Idade Lote	DACT	Total dia (x1000)	Id MD Inc	Luz
128	86	< 8	5-19		> 58		144-161	< 7	
21	76	< 8	< 5		58-63		103-259	< 7	
69	75	< 8	5-19	< 33	> 58		< 144	< 7	
33	69		> 19		> 58			< 7	726
61	67	< 7			> 58		< 285	> 7	
394	65	< 8	5-19		> 58		161-259	< 7	
164	54	< 8	< 19		> 58	< 14	> 259	< 7	

Nro. Reg.=Número de Registros que atendem à regra.

Pode ser observado na Figura 2, que a cada nível um atributo é identificado mais importante para gerar 2 grupos de dados semelhantes entre si e diferentes entre grupos. A AD gerada tem 9 níveis de profundidade e a partir daí, os grupos gerados não atendem às configurações disponibilizadas ao algoritmo, e dessa forma a classificação foi encerrada.

A Tabela 4 gerada é a descrição da AD da Figura 2.

Experimento NEG

Na Tabela 5 são apresentados os resultados da seleção de variáveis do software SAS-EM, para o experimento NEG.

Para esse experimento, foram rejeitados os atributos Temperatura acima do conforto térmico, Número de dias consecutivos em desconforto térmicos e AT.

No experimento NEG, buscou-se identificar os padrões de incubação dos lotes que tiveram desempenho mais desfavorável.

Os dados da Tabela 5 indicam que o atributo “Idade do Ovo” e “Idade do Lote” são os de maior impacto nos resultados menos favoráveis. Outros atributos também influenciam, mas são de menor impacto, conforme o índice *Importance*.

A temperatura ambiente (mínima e máxima) no dia da postura, que poderia ser uma causa desfavorável para incubação, revelou-se um atributo que influencia pouco os resultados.

Os atributos Temperatura acima do conforto térmico, Dias com temperatura acima do conforto térmico e AT, foram rejeitados para análise por apresentarem baixo valor de *Importance*.

A Árvore de Decisão gerada é apresentada na Figura 3.

Na Figura 3 pode ser observado que o atributo "Idade Ovo" foi o mais importante para iniciar o agrupamento dos dados (*split* inicial), seguido de "Idade Lote" e que, em níveis mais baixos da AD, são utilizados novamente para classificar e criar grupos ainda mais homogêneos. Na regra 3, no terceiro nível da AD, a classificação é encerrada com a criação de grupo de dados similar (folha ou *node*). Em outros níveis da AD são utilizados outros atributos para classificação.

A Tabela 6 apresenta de forma descritiva a AD da Figura 3.

Foram identificadas 8 regras com acurácia entre 52 e 88%, como apresentado na Tabela 6.

Tabela 5 – Atributos selecionados pelo módulo Seleção de Variáveis do SAS-EM para o experimento NEG.

Atributos	Hy-Line W-36	
	Importance	Ordem
Idade ovo – IO	1,0000	1
Idade do lote – Idade lote	0,7321	2
Número de ovos incubados do lote – Ovo Inc	0,3375	3
Total de ovos incubados no dia – Total dia	0,3306	4
Idade Média Incubação - Id MD Inc (dias)	0,1958	5
Temperatura Mínima (^{o}C) – Min	0,1939	6
Temperatura Máxima (^{o}C) - Max	0,1609	7
Luz (minutos/dia)	0,1326	8
Número de lotes incubados dia – Nr lote inc	0,0996	9
Temperatura acima do conforto térmico (S/N)		
Dias com temperatura acima do conforto térmico-DACT		
Amplitude Térmica		

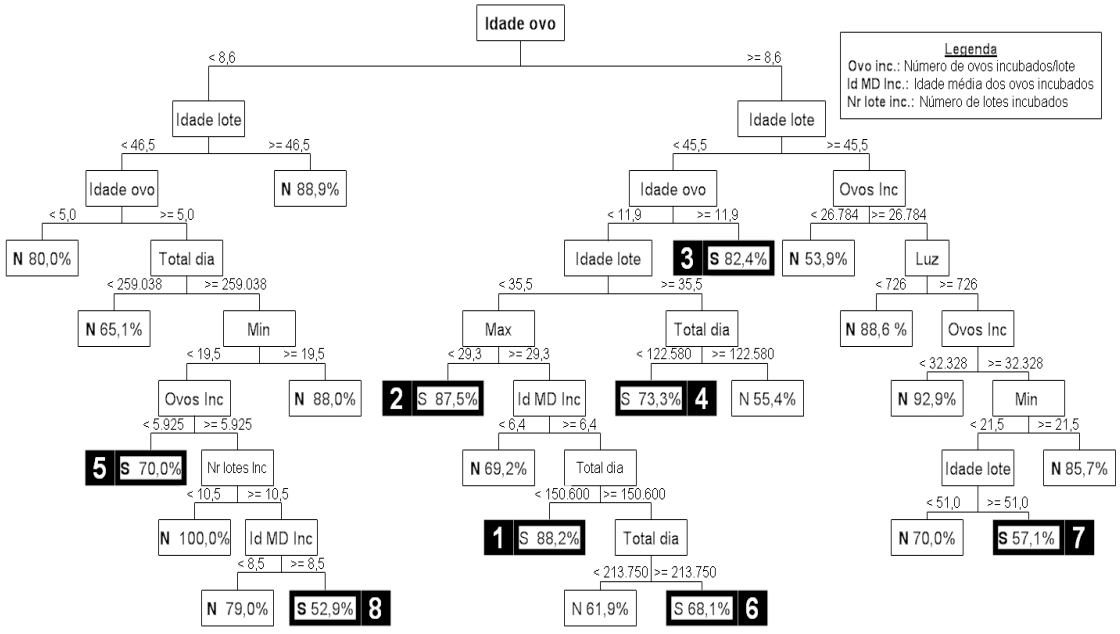

Figura 3 – Árvore de Decisão adaptada, experimento NEG.

Tabela 6 – Regras de interesse geradas pelo SAS-EM do Experimento NEG para a linhagem Hy-Line W-36.

Nro.Reg.	Acurácia (%)	IO	Temperatura		Ovo Inc (x1000)	Idade lote	Total dia (x1000)	Id MD Inc	Luz	Nr Lote inc
			Min	Max						
17	88	8-11		> 29		< 35	<150	> 6		
32	87	8-11		< 29		< 35				
68	82	> 11				< 45				
30	73	8-11				35-45	< 122			
10	70	5-8	> 19		< 5	< 46	> 259			
47	68	8-11		> 29		< 35	> 213	> 6		
21	57	> 8	< 21		> 32	> 51			> 726	
17	52	5-8	> 19		> 5	< 46	> 259			> 10

Nro.Reg.=Número de Registros que atendem a regra.

Nos dois experimentos foram identificados conhecimentos de interesse. Nas Tabelas 3 e 5 pode ser observado que os atributos "Idade do Lote" e "Idade do Ovo" foram os atributos mais importantes (valores altos de *Importance*). Mas foram de resultado oposto, sendo que para se obter resultados positivos de incubação, os ovos devem ser utilizados com até 7 dias de estocagem e ovos mais velhos que 8 dias foram observados nos resultados menos favoráveis de incubação.

Experimento POS

Pode ser observado, na Tabela 3, que os atributos Idade do Lote, Número de lotes incubados, Idade média dos ovos incubados, Idade do ovo, Luz e Número de ovos incubados, são importantes para identificação de padrões nos dados.

Neste experimento foram classificadas como "S" as incubações cujos resultados eram, no mínimo, 4,4 pontos porcentuais ou mais, superiores aos valores esperados.

Na Tabela 3 pode ser observado que o atributo de maior importância, "Idade do Lote", é próximo de 4 vezes mais importante que o atributo subsequente, confirmando seu potencial.

Aves da linhagem Hy-Line W-36 com mais de 58 semanas de idade (ciclo final de produção), cujos ovos foram estocados por menos de 7 dias, apresentaram produtividade superior. Vieira et al. (2005), em experimentos com linhagens de frango de corte, encontraram resultados semelhantes.

Outros atributos foram incluídos em algumas regras, mas com comportamento variável.

Experimento NEG

Resultados de incubação com 2,0 pontos porcentuais de diferença dos valores esperados, ou menos, foram considerados resultados menos favoráveis.

Na Tabela 6 indica-se que aves jovens, com idade inferior a 35 semanas de idade, cujos ovos foram produzidos em dia com temperatura ambiente máxima acima de 29ºC, e que foram estocados entre 8 e 11 dias, apresentaram resultados menos favoráveis de incubação. Outros atributos também contribuíram para o desempenho menos favoráveis, apresentados na Tabela 5.

Simulação de resultados

Para facilitar o entendimento do conhecimento identificado, é proposto o cenário a seguir que simula um possível ganho para a empresa.

Supondo uma demanda de 7.000 pintinhas vendáveis, utilizando-se o valor de referência de produtividade da empresa, 42% de pintinhas nascidas vendáveis, seria necessário incubar aproximadamente 16.700 ovos férteis. Baseado no conhecimento de uma das regras identificadas, em que 46,4% passaria a ser a produtividade, poderiam ser incubados aproximadamente 15.100 ovos, com uma redução de 1.600 ovos, ou 9,65% a menos.

CONCLUSÕES

Esta pesquisa nos permite concluir que a técnica Árvore de Decisão identifica padrões de produção de dados de incubação de ovos de matrizes de postura. Os dados de gerenciamento das incubações foram suficientes para análise. Considerável esforço foi realizado para o pré-processamento dos dados, próximo de 90% do tempo do projeto. As regras geradas pela técnica foram interpretadas e validadas pelo especialista de incubação. O conhecimento gerado pelas regras sugere

ajustes nos programas de incubação, potencializando a produção de fêmeas vendáveis e melhorando os resultados do negócio.

AGRADECIMENTOS

Às empresas Hy-Line do Brasil Ltda, em especial ao Dr. Antonio Carlos Paraguassú , SAS Institute Brasil, a Andréa Szyfer, e o IAC-Campinas, pelo apoio no desenvolvimento desta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

ELIBOL, O.; PEAK, S. D.; BRAKE, J. Effect of flock age, length of egg storage, and frequency of turning during storage on hatchability of broiler hatching eggs. **Poultry Science**, v. 81, n. 7, p. 945-950, 2002.

FAYYAD, U.M.; PIATETSY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery. In: FAYYAD, U.M.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Eds.). **Advancews in knowledge discovery and data mining**. Menlo Park: AAAI/MIT, 1996. p.1-24.

FREITAS, H.J.de; COTTA, T.deB.; OLIVEIRA, A.I.G. de; GEWHER, C.E. Avaliação de programas de iluminação sobre o desempenho zootécnico de poedeira leves. **Ciência e Agrotecnologia**, Lavras, v. 29, n. 2, p. 424-428, mar./abr., 2005.

KIRCHNER, K.; TÖLLE, K.H.; KRIETER, J. Decision tree technique to pig farming datasets. **Livestock Production Science**, Amsterdam, n.90, p.191-200, 2004.

OLIVEIRA, A.C.S.de; SOUZA, A.A.de; LACERDA, W. S.; GONSALVES, L.R.Aplicação de redes neurais artificiais na previsão da produção de álcool. **Ciência e Agrotecnologia**, Lavras, v. 34, n. 2, p. 279-284, mar./abr., 2010.

ROUSH, W.B.; KOCHERA, Y.; KIRBY, T.L.; CRAVENER, T.; WIDEMAN JUNIOR, R.F. Probabilistic neural network prediction of ascites in broiler based on minimally invasive physiological factors. **Poultry Science,** v.76, p.1513-1516, 1996.

SAS INSTITUTE. **Enterprise miner**: help. Version 4.3. Cary, 2005.

VIEIRA, S.L.; ALAMEIDA, J.G.; LIMA, A.R.; CONDE ORA, O.A.R. Hatching distribution of eggs varying in weight ans breeder age. **Brazilian Journal of Poultry Science,** São Paulo, p.73-78, 2005.

VIEIRA, S.L.; MORAN, E.T. Comparison of eggs and chicks from broiler breeders of extremely different ages. **Journal of Applied Poultry Research**, London, v.7, p.372-376, 1998.