UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

FELIPE ADI HANKE

ANÁLISE DA PRODUTIVIDADE DE BOVINOS LEITEIROS POR MEIO DA APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2019

FELIPE ADI HANKE

ANÁLISE DA PRODUTIVIDADE DE BOVINOS LEITEIROS POR MEIO DA APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Computação".

Orientador: Prof. Dr. Alan Gavioli

MEDIANEIRA

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

ANÁLISE DA PRODUTIVIDADE DE BOVINOS LEITEIROS POR MEIO DA APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

Por

FELIPE ADI HANKE

Este Trabalho de Conclusão de Curso foi apresentado às 13:00 do dia 19 de novembro de 2018 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Alan Gavioli
UTFPR - Câmpus Medianeira

Prof. Evando Carlos Pessini
UTFPR - Câmpus Medianeira

Prof. Cesar Angonese
UTFPR - Câmpus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

Hanke, Felipe Adi. ANÁLISE DA PRODUTIVIDADE DE BOVINOS LEITEIROS POR MEIO DA APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS. 62 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

A tecnologia tornou-se uma grande aliada dos produtores de bovinos leiteiros, mas poucos produtores conseguiram obter tais tecnologias, continuando a trabalhar com o conhecimento empírico acumulado durante os anos. No trabalho desenvolvido, foram abordadas as técnicas de mineração de dados, com a finalidade de analisar dados de bovinos leiteiros e auxiliar a empresa a utilizá-los, tentando identificar variáveis que podem exercer influência na produtividade de leite, com intuito de obter conhecimento possivelmente novo e útil para os proprietários da fazenda considerada. Foram analisados os dados de uma empresa de produção leiteira de Céu Azul, Paraná.

Palavras-chave: Bovinos leiteiros, mineração de dados, produtividade.

ABSTRACT

Hanke, Felipe Adi. ANALYSIS OF THE PRODUCTIVITY OF DAIRY CATTLE THROUGH THE APPLICATION OF DATA MINING TECHNIQUES. 62 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

The technology has become a major ally of dairy cattle producers, but just some producers have been able to obtain these technologies, continuing to work with empirical knowledge accumulated over the years. In the developed work, data mining techniques will be approached in order to analyze dairy cattle data and help the producers to improve productivity and/or improve the use of generated data, in order to obtain possibly new and useful knowledge to the owners of the farm considered. It will be analyzed using the data of a dairy company in Céu Azul, Paraná.

Keywords: Dairy cattle, data mining, productivity.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

IBGE	Instituto Brasileiro de Geografia e Estatística
MDIC	Ministério do Desenvolvimento, Indústria e Comércio Exterior
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
KDD	Knowledge Discovery in Databases
ITU	Índice Térmico de Umidade
WEKA	Waikato Environment for Knowledge Analysis
GPL	General Public License
API	Application Programming Interface
CCS	Contagem de Células Somáticas
RFID	Identificação por Rádio Frequência

SUMÁRIO

1 INTRODUÇÃO

De acordo com uma pesquisa do Instituto Brasileiro de Geografia e Estatística (IBGE), publicada em setembro de 2016, o Paraná se manteve em segundo lugar na produção leiteira, tendo um aumento de 1,5% na produtividade em relação ao último ano. A pesquisa aponta que nos últimos 10 anos o setor teve um acréscimo de 75% alcançando o crescimento de 2 bilhões de litros de leite neste período, mesmo contando com a redução na produção de 2,8% no ano de 2016 comparado com a produtividade do ano anterior (2015) (Agência de notícias do Paraná, 2017).

Assim como o Paraná, os outros estados da região sul do Brasil também tiveram um aumento na produtividade, colocando a região como maior produtora do país com o volume de 12,45 bilhões de litros de leite produzidos, sendo 1,5% mais produtivo que o ano anterior. Mesmo com o aumento da região sul, o estado de Minas Gerais ainda detém o título de maior produtor de leite do país, com um volume de 8,87 milhões de litros de leite, volume este que teve uma queda de 1,9% comparado com o ano de 2015.

No ano de 2016 muitas áreas do país sofreram com queda de produtividade, devido a fatores externos que prejudicam diretamente a classe produtora. O país estava passando por uma crise, o que fez gerar inflação e aumento de alguns produtos necessários para a produção de leite, desta forma a inflação atacou diretamente a população diminuindo o seu poder de compra. Os fatores climáticos também ajudaram, em alguns lugares com seca e outros com excesso de chuva. A região sul foi a menos afetada e com isso conseguiu um crescimento notável em relação às outras regiões.

A receita de importação e exportação de lácteos no Brasil teve um grande déficit no ano de 2017, de acordo com os dados divulgados pelo Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC). A importação caiu 15% em relação a 2016, já nas exportações a queda foi grande, chegando aos 34% em relação ao ano anterior (Universo Online, 2018). Dados preocupantes, visto que países vizinhos conseguiram crescer muito seus mercados neste mesmo setor.

Há alguns anos, os produtores observaram a diferença que a tecnologia pode fazer em suas fazendas. Com isso, o investimento em tecnologia aumentou muito, fator que

incentiva os pesquisadores a procurar novas formas de inovar o mercado com a adoção de melhores técnicas. Uma das inovações aplicadas é a mineração de dados, que pode proporcionar resultados promissores no que tange à melhoria da tomada de decisões estratégicas para as empresas. Algumas pesquisas relevantes já foram publicadas, por exemplo uma que confirma que o conforto térmico dos bovinos leiteiros é de grande importância, pois influencia significativamente na produtividade e no manejo daqueles animais com um conforto considerado alto (temperatura retal abaixo de 38,8^{o}C e frequência respiratória abaixo de 56 mov/min (PERISSINOTTO et al., 2009)). Outro exemplo significativo foi a criação de um novo método de observação do estro (período em que o animal está no cio), algo que até então era primitivo (feito por meio de simples análise visual por produtores mais experientes). Mas com as técnicas de mineração de dados, pode-se determinar o estro animal apenas com a movimentação dos animais, devido à movimentação de cada vaca, apenas utilizando um pedômetro e um software que monitora e guarda os dados.

A Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) é uma forte aliada dos produtores no avanço tecnológico, a empresa presta grandes serviços para a comunidade, pois leva até os produtores novidades, estas que ajudam o Brasil de forma geral.

1.1 OBJETIVO GERAL

Analisar, por meio da aplicação de técnicas de classificação de mineração de dados, a influência de características do gado leiteiro na produtividade de leite de uma fazenda no estado do Paraná.

1.2 OBJETIVOS ESPECÍFICOS

- Constituir um banco de dados com informações correspondentes à criação de gado leiteiro de uma fazenda do estado do Paraná;
- Aplicar técnicas de mineração de dados sobre o banco de dados construído, para identificar variáveis que exerçam influência sobre a produtividade de leite;

- Analisar os resultados obtidos pela aplicação de mineração de dados, com intuito de obter conhecimento possivelmente novo e útil para os proprietários da fazenda considerada.

1.3 JUSTIFICATIVA

O avanço da produção de leite no Brasil, gera necessidades de progressos tecnológicos e mais informações úteis para os produtores leiteiros. Visto que pode-se gerar um grande volume de dados em uma fazenda, aproveitá-los da melhor forma possível é importante.

Com o desenvolvimento deste trabalho, espera-se contribuir para a melhoria dessa atividade produtiva, com a possível obtenção de conhecimento novo e útil para os produtores. Neste sentido, pretende-se analisar os atributos relevantes no aumento da produtividade de leite.

2 REFERENCIAL TEÓRICO

Neste capítulo, será apresentado o que é mineração de dados, assim como as técnicas que podem vir a ser utilizadas para a realização deste trabalho. O conceito de produtividade de gado leiteiro e pesquisas que relacionam mineração de dados e produtividade de leite.

2.1 KDD

Segundo Fayyad et al. (1996), há um grande volume de dados sendo gerado diariamente, mas pouco aproveitado. Portanto, vê-se uma necessidade de usar estes dados de uma maneira mais útil, transformando dados em informações. Para fazer o melhor aproveitamento dos dados, sugere-se o processo chamado, *Knowledge Discovery in Databases* (KDD).

Este processo é dividido em várias fases, sendo elas: Coleta de dados, pré-processamento, transformação, mineração de dados, avaliação e interpretação. Na Figura 1, pode-se ver como este processamento ocorre (FAYYAD et al., 1996).

- **Seleção de dados:** Um processo que pode ser muito trabalhoso, pois os dados em bancos de dados transacionais podem não estar tratados para se fazer a análise de dados. Mas muito importante, pois a escolha do conjunto de dados que faz com que o resultado final seja conclusivo. A escolha dos dados depende muito dos resultados esperados, portanto a ajuda de um especialista da área do problema é importante. (FAYYAD et al., 1996).
- **Pré-Processamento:** Esta fase é responsável por padronizar os dados, excluindo ruídos, valores desconhecidos, atributos de baixo valor preditivo, entre outros (BATISTA, 2003);
- **Transformação dos dados:** Depois de realizado o pré-processamento, os dados precisam ser transformados, de acordo com o tipo de algoritmo que será utilizado. Alguns tipos comuns de transformação são: normalização (consiste em padronizar os valores dos atributos em uma única escala), discretização de atributos quantitativos (alguns

algoritmos trabalham apenas com um tipo de dado, sendo necessário a transformação para o padrão correto), suavização (remove atributos incorretos dos dados) (BATISTA, 2003);

- **Mineração de dados:** A fase de mineração de dados é responsável pela escolha do algoritmo certo para a resolução do problema. Para isso o estudo realizado sobre a natureza do problema é muito importante, pois assim a escolha do algoritmo será fácil e assertiva. Não existe um algoritmo ótimo para todas as situações, em alguns casos onde tem-se volumes de dados maiores, existe a necessidade de combinar resultados entre vários algoritmos, sendo melhor que escolher um único algoritmo (BATISTA, 2003; FAYYAD et al., 1996);
- **Avaliação e interpretação de resultados:** Nesta fase é onde o analista de dados, analisa se a maneira em que o processo foi realizado, foi de maneira correta ou não. Analisando os resultados de acordo com as taxas de erro, tempo de CPU e complexidade do modelo. O especialista na natureza do problema, analisa os resultados com base no próprio conhecimento. E fechando o processo, o usuário faz seu julgamento sobre a aplicabilidade dos resultados obtidos (BATISTA, 2003).

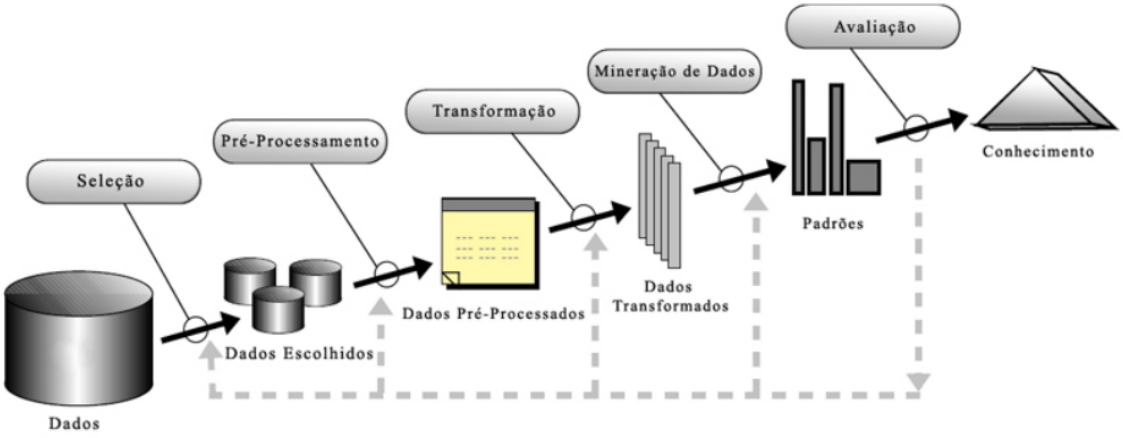

Figura 1 – Figura representativa do processo KDD

Fonte: Camilo e Silva (2009)

2.2 MINERAÇÃO DE DADOS

A mineração de dados (*data mining*, em inglês) é o estudo que busca extrair informações relevantes através de coleta, análise e processamento de um grande volume de dados. Portanto, a mineração de dados é um termo amplo que assim como a descoberta de conhecimento em inteligência artificial, busca descobrir automaticamente regras e modelos estatísticos a partir dos dados (AGGARWAL, 2015; CARVALHO, 2001).

Praticamente todos os atuais sistemas automatizados geram um grande volume de dados, que são importantes para fins de análise e diagnóstico. Este grande volume de dados está ligado diretamente aos avanços tecnológicos, portanto é natural usar dos mesmos para extrair conhecimento novo e útil. Neste ponto, a mineração de dados se torna importante, pois muitas vezes permite extrair informação relevante de grandes volumes de dados (AGGARWAL, 2015).

2.3 TÉCNICAS DE MINERAÇÃO DE DADOS

Nesta seção serão abordadas as principais técnicas de mineração de dados, associação, classificação, agrupamento e regressão, respectivamente. Serão apresentados os conceitos fundamentais e suas principais aplicações.

2.3.1 Associação

As regras de associação foram originalmente desenvolvidas para melhorar o marketing usado pelas empresas. Estudando o comportamento dos clientes na compra dos produtos, fazendo uma análise contextual usando o algoritmo de regras de associação. Itens que são associados com grande frequência podem ser melhor dispostos em lojas de varejo, site de comércio eletrônico ou em outros exemplos que têm o mesmo intuito, a melhoria do marketing

(YE, 2013).

A mineração de regras de associação proporciona conhecimentos valiosos na avaliação de analogias significativas que podem ser encontradas em grandes bancos de dados. Formalmente, uma regra de associação é definida pela expressão $X \rightarrow Y$ (lê-se: se X então Y), em que X e Y são conjuntos disjuntos de itens (WITTEN et al., 2011; MICHALSKI et al., 1998).

Dois tipos de medidas são muito importantes nas regras de associação, sendo elas as medidas de suporte e de confiança. A medida de suporte se refere ao grau de relevância que aparece nos dados, por exemplo, usando uma regra onde $X \rightarrow Y$, a medida de suporte irá mostrar a porcentagem onde os itens X e Y aparecem na base de dados, mostrando assim sua relevância. A medida de confiança indica a porcentagem de transações que contêm os itens de X e Y juntos, sobre o total de transações que possuem os itens de X (sendo X o lado esquerdo da regra) Na Figura 2, esta é a representação de três regras de associação, onde na primeira regra pode-se analisar um produto comparado a outro (Leite $\rightarrow$ Pão), na segunda regra pode-se analisar dois produtos comparado a um produto (Presunto, Queijo $\rightarrow$ Pão) e na terceira regra pode-se analisar um produto comparado a outros dois produtos (Carvão $\rightarrow$ Carne, Cerveja) (WITTEN et al., 2011; MICHALSKI et al., 1998).

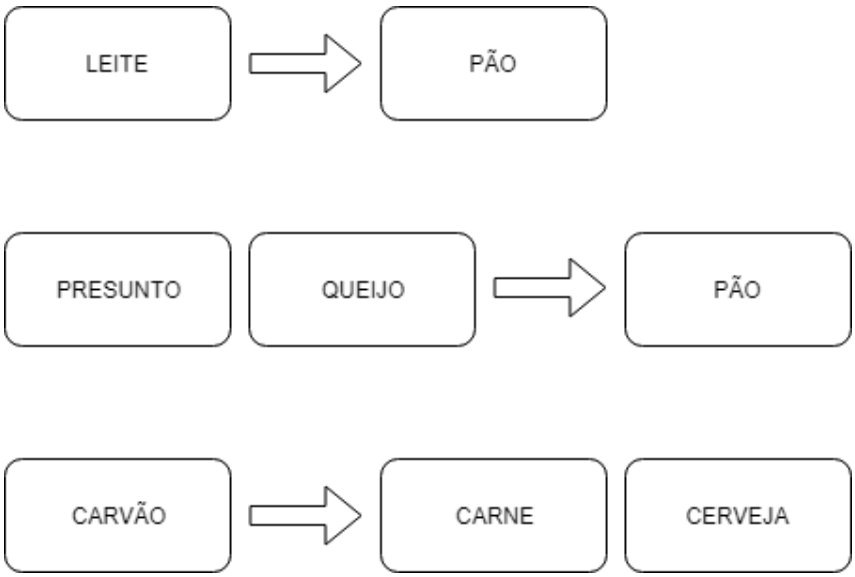

Figura 2 – Exemplos de regras de associação

Fonte: Autoria própria

Com o intuito de encontrar regras de associação para grandes bancos de dados, foram sugeridos diversos algoritmos. A maior parte, baseiam-se em atributos categóricos manipulando valores binários, reconhecendo correlações entre operações (MICHALSKI et al., 1998; AGRAWAL; SRIKANT, 1994).

2.3.2 Agrupamento

O agrupamento é uma tarefa de mineração não supervisionada, em que o analista não fornece informação para a ferramenta de treinamento. Tem como objetivo formar grupos que contenham instâncias que apresentem semelhanças entre si. Neste sentido, algoritmos de agrupamento buscam garantir elevada similaridade dentro de cada grupo e dissimilaridade entre grupos. Ela não se preocupa em prever valores ou classificar, apenas separar em grupos onde os dados sejam parecidos. Isto é ilustrado na Figura 3, em que pode-se ver três grupos distintos, definidos por um algoritmo de agrupamento (WITTEN et al., 2011; AGGARWAL, 2015; CAMILO; SILVA, 2009; HAN; KAMBER, 2005; PAWET, 2015).

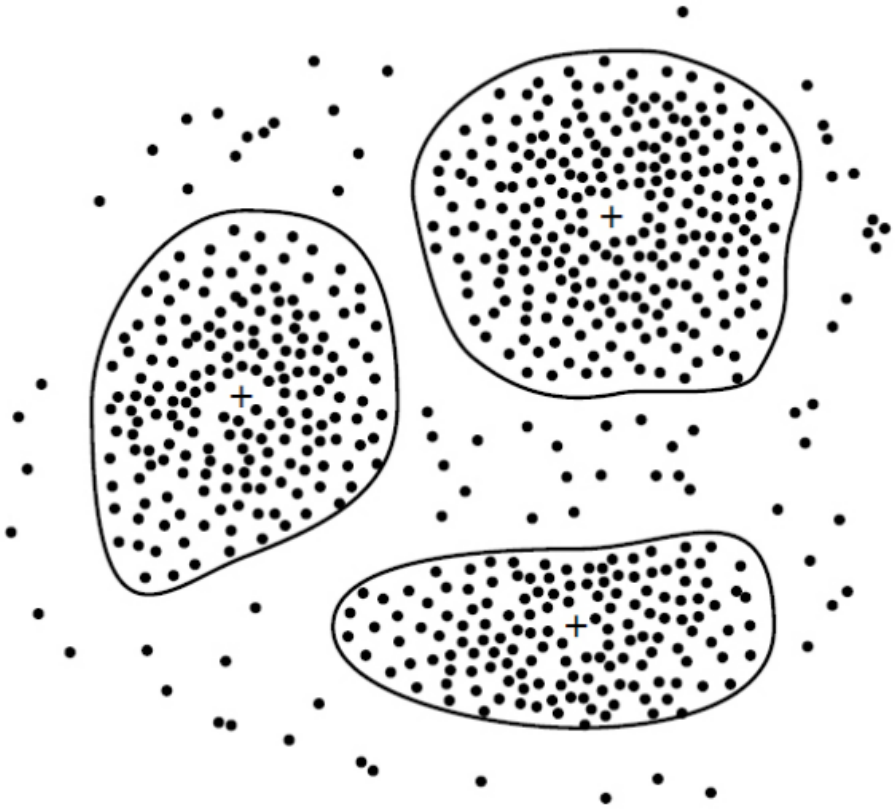

Figura 3 – Registros separados em 3 Clusters
Fonte: Han e Kamber (2005)

Uma forma simples de exemplificar é pensar em uma criança brincando com a caixa de botões de roupas de sua avó. Dentro da caixa existem os mais variados tipos de botões, sendo eles grandes, pequenos, com um ou mais furos, com formatos diferentes ou cores diferentes.

Podem-se agrupar os botões de diversas formas diferentes, depende a necessidade do objetivo que deseja alcançar. No exemplo da Figura 4 apresentam conjuntos de objetos heterogêneos, onde classifica-se de três formas distintas, separando por cor, forma ou a junção de forma e cor (BROWN, 2014).

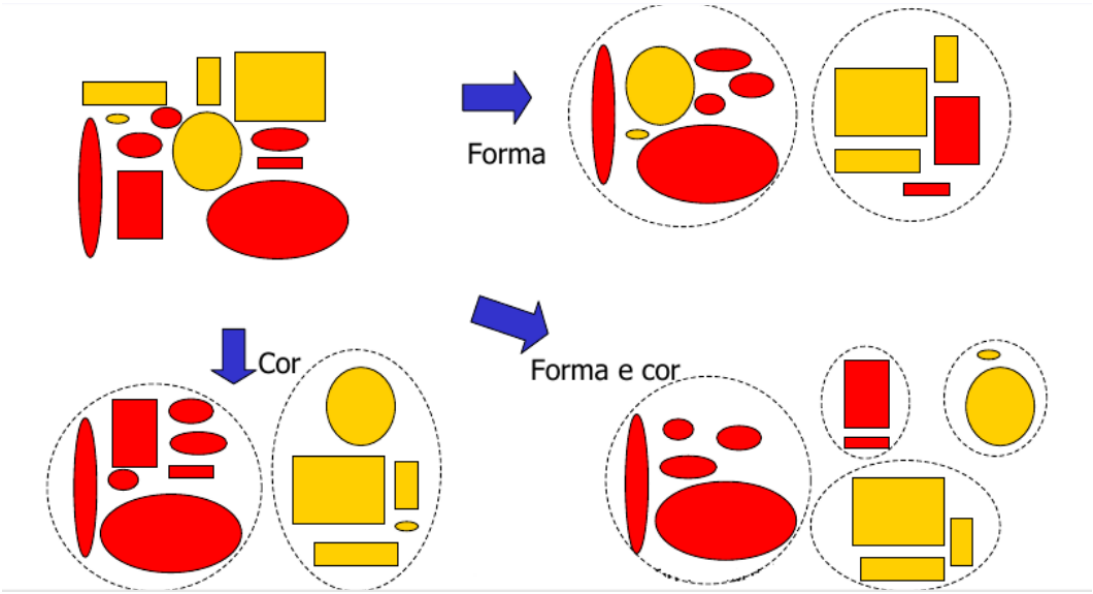

Figura 4 – Exemplo de possíveis formas de separar grupos

Fonte: FACELI et al. (2011)

O agrupamento tem uma grande utilidade no mundo real. Pode-se citar algumas: agrupa clientes de várias maneiras diferentes com base nos atributos, descobre os lugares mais adequados para cada tipo de negócio, cria listas de produtos para cada tipo de cliente, é usado para detecção de invasão em redes como também em transações bancárias para identificar fraudes (CAMILO; SILVA, 2009; PAWET, 2015).

Para ser realizado o bom uso dos dados de entrada, eles passam por uma seleção de recursos, que tem como finalidade remover os atributos que não sejam efetivamente úteis para um agrupamento adequado. O agrupamento, por ser uma tarefa não supervisionada, gera uma maior complexidade na escolha de recursos, pois faltam parâmetros externos de validação, ou seja, faltam informações dadas pelo analista como corretas para o algoritmo se fundamentar (AGGARWAL, 2015).

São vários os algoritmos que conseguem executar o agrupamento, mas dentre eles um se destaca, considerado clássico na resolução de problemas de agrupamento, conhecido como K-Médias. O algoritmo K-Médias tem uma funcionalidade baseada no aprendizado de máquina, usando técnicas de distâncias para mover os K's centrais para o lugar mais correto. O algoritmo é implementado em quatro passos: primeiro passo, define-se K objetos centrais de K

grupos, aleatoriamente; segundo passo, os pontos são atribuídos aos K's centrais mais próximos; terceiro passo, após o segundo passo são formados K's grupos, após formados são recalculados os K's centrais para que estes estejam no meio de cada grupo formado; quarto passo, o segundo e o terceiro passos são repetidos até que os K's centrais não mudem de posição, com isso obtém-se o resultado de saída do algoritmo, os K's grupos formados (WITTEN et al., 2011; AGGARWAL, 2015; HAN; KAMBER, 2005).

Devido a aleatoriedade do K central, os resultados dificilmente são confiáveis caso o algoritmo seja executado uma única vez, portanto o agrupamento é feito várias vezes até que o algoritmo encontre o melhor resultado final, sendo constatado o melhor resultado assim que encontrado a menor distância quadrada total (WITTEN et al., 2011).

2.3.3 Classificação

A classificação pode ajudar a resolver muitos problemas reais comuns, problemas que quando resolvidos ajudam de maneira significativa as áreas em que são exploradas. Dois exemplos básicos, a quantidade de crédito que um banco pode ceder a uma pessoa desconhecida pela instituição, e um pesquisador da área médica analisando dados de uma doença para apontar os possíveis tratamentos. Cada um desses exemplos podem ser resolvidos com o método de classificação, podendo com um classificador eficiente apontar as melhores escolhas, tanto no caso do empréstimo bancário, onde o método pode apontar para quais pessoas é seguro ou arriscado liberar dinheiro, como para o médico pesquisador, que consegue encontrar de forma rápida e simples os tratamentos certos para cada paciente (HAN; KAMBER, 2005).

No aprendizado supervisionado o analista de dados fornece respostas corretas, ajudando a ferramenta de aprendizado, conseguindo com isso melhores resultados. No aprendizado não supervisionado, o analista não fornece nenhum tipo de ajuda ao algoritmo, a ferramenta usa normalmente fórmulas matemáticas ou regularidades encontradas nos dados (HAN; KAMBER, 2005).

Sendo uma das funções mais comuns, a classificação tem como objetivo distinguir qual classe um determinado atributo condiz. A classificação tem duas etapas distintas, na primeira etapa se constrói o modelo de classificação com um conjunto de classes pré determinadas, cada registro conhecido tem uma classe pré-definida, estipulada pelo valor do atributo-alvo. O conjunto criado na construção do modelo é denominado como conjunto de treinamento. O

modelo construído pode ser representado por regras de classificação, árvores de decisão ou fórmulas matemáticas. A segunda etapa tem a função de utilizar o modelo criado na primeira etapa, fazendo a validação do conjunto de treinamento com um conjunto de testes diferente do conjunto de treinamento. Além de fazer a classificação de futuros registros ou valores desconhecidos dentro do conjunto (AGGARWAL, 2015; HAN; KAMBER, 2005; CAMILO; SILVA, 2009). Na Tabela 5 [1], pode-se ver um exemplo de tuplas com classes preditivas (Céu, Temperatura, Umidade, Vento) e o atributo-alvo (Jogar Tênis).

Céu	Temperatura	Umidade	Vento	JogarTênis
Ensolarado	Quente	Alta	Fraco	NÃO
Ensolarado	Quente	Alta	Forte	NÃO
Nublado	Quente	Alta	Fraco	SIM
Chuvoso	Boa	Alta	Fraco	SIM
Chuvoso	Fria	Normal	Fraco	SIM
Chuvoso	Fria	Normal	Forte	NÃO
Nublado	Fria	Normal	Forte	SIM
Ensolarado	Boa	Alta	Fraco	NÃO
Ensolarado	Fria	Normal	Fraco	SIM
Chuvoso	Boa	Normal	Fraco	SIM
Ensolarado	Boa	Normal	Forte	SIM
Nublado	Boa	Alta	Forte	SIM
Nublado	Quente	Normal	Fraco	SIM
Chuvoso	Boa	Alta	Forte	NÃO

Figura 5 – Conjunto de treinamento, com os atributos preditivos e atributo-alvo

Fonte: Autoria própria

Costumeiramente é citada como aprendizado supervisionado, pois faz referência direta com um professor supervisionando seus alunos com objetivo de aprender determinado conteúdo. Portanto, é primordial que os dados que são fornecidos como dados de treinamento (dados que foram divididos em suas respectivas classes) sejam bons exemplos. Usando um volume de dados de teste, eles irão usar os dados de treinamento para se espelhar e assim atribuir classes às instâncias novas. (AGGARWAL, 2015).

Existem diversos algoritmos para classificação de dados. Um dos mais simples e utilizados é a árvore de decisão (HAN; KAMBER, 2005). A árvore de decisão é muito similar a um fluxograma, em que pode-se analisar os nós e as folhas da seguinte maneira: cada nó (não folha) representa um teste feito ao valor de um atributo de um elemento conjunto de dados,

[1] Retirada da base de dados do WEKA

os possíveis valores desse atributo estão apresentados nos ramos, e com cada folha obtém-se o resultado sobre a classe na qual deve ser incluído o elemento considerado (WITTEN et al., 2011; CAMILO; SILVA, 2009; HAN; KAMBER, 2005).

Na Figura 6, mostra-se a árvore de decisão para possíveis compradores de computador. Nos nós são representados os testes de cada atributo, nos ramos as possíveis respostas e nas folhas as classes. Alguns algoritmos produzem apenas árvores binárias, enquanto outros podem gerar não binárias (HAN; KAMBER, 2005). Com facilidade consegue-se converter as árvores de decisão em regras de classificação, por exemplo, seguindo a Figura 6.

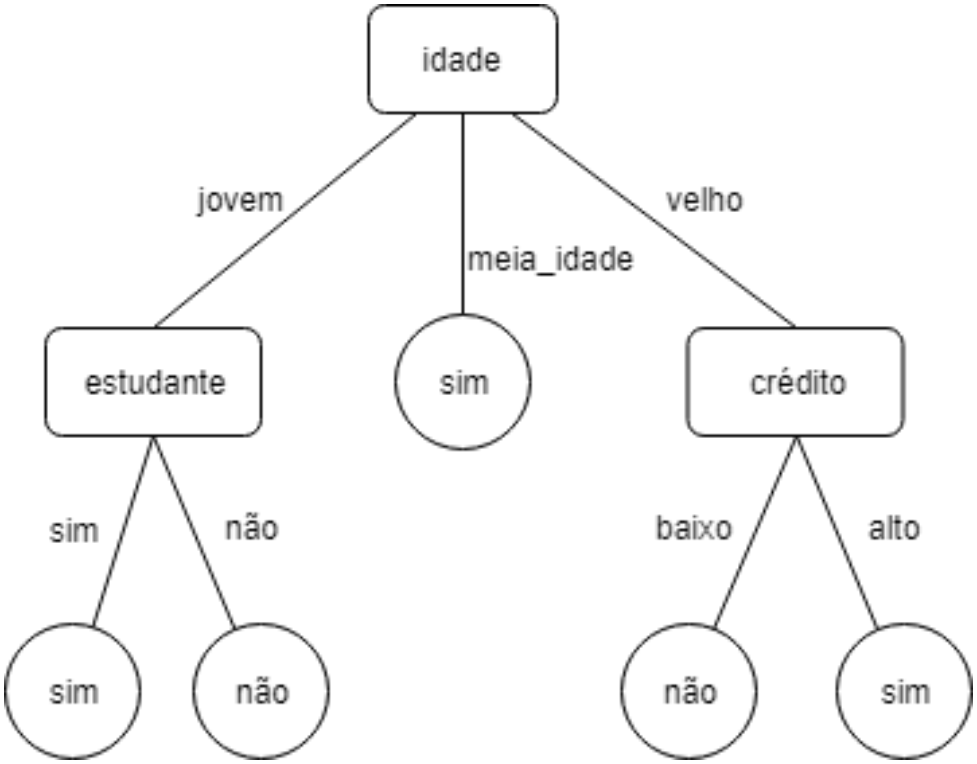

Figura 6 – Árvore de decisão
Fonte: Adaptado de (HAN; KAMBER, 2005)

IF (idade = jovem **AND** estudante = sim) **OR**
(idade = meia_idade) **OR**
(idade = velho **AND** crédito = alto)
THEN ComprarComputador = SIM

Existem vários motivos para as árvore de decisão serem populares, como por exemplo: a simplicidade de resolver os problemas, a facilidade para um humano entender o que está acontecendo intuitivamente, o problema de dimensionalidade é resolvido com facilidade, as etapas de aprendizagem e classificação são rápidas e simples. Todos esses prós combinam com a boa precisão que os classificadores das árvores de decisão tem, embora os resultados

estão ligados diretamente aos dados fornecidos, portanto bons dados, geram bons resultados. Os algoritmos foram usados em muitas áreas com sucesso, desde áreas de Ciências Exatas até mesmo Ciências Biológicas, mostrando sua efetividade e importância (CAMILO; SILVA, 2009; HAN; KAMBER, 2005).

2.3.4 Regressão

O modelo de regressão tem similaridade com o modelo de classificação, tendo diferença nos tipos de atributos que são classificados: no modelo de regressão, o trabalho é realizado sobre atributos quantitativos, ao invés de atributos discretos (CAMILO; SILVA, 2009; PAWET, 2015).

A técnica de regressão é importante por fazer previsões numéricas. Pode-se prever a quantidade de produção, valores de objetos na compra/venda e/ou a previsão de próximos salários. As técnicas de regressão podem ser usadas quando existe uma relação, entre uma ou mais variáveis preditoras com a classe que pretende-se predizer. Assim como na tarefa de classificação, a variável que se quer prever pode ser chamada de atributo-alvo. Com os atributos descritos nas variáveis preditoras quer-se encontrar os valores do atributo-alvo (PAWET, 2015; LAROSE; LAROSE, 2015; YE, 2013; HAN; KAMBER, 2005).

Os problemas em regressão podem ser lineares ou não lineares. Para problemas não lineares, pode-se aplicar a transformação de variáveis, de modo que o problema passe a ser um problema linear. Quando tem-se um problema onde a variável y tem um valor linear ao de x, pode-se empregar regressão linear. Exemplo: Na Figura 7, pode-se ver um problema linear, onde avalia-se os anos de experiência em relação aos salários mínimos (valor mínimo, definido por lei, para que uma pessoa mantenha sua sobrevivência) (CAMILO; SILVA, 2009; PAWET, 2015).

2.4 PRODUÇÃO PECUÁRIA

A produção pecuária é o processo de criação de animais domesticados, desde a produção do alimento até a venda do animal ou de seus produtos. Quando refere-se à pecuária,

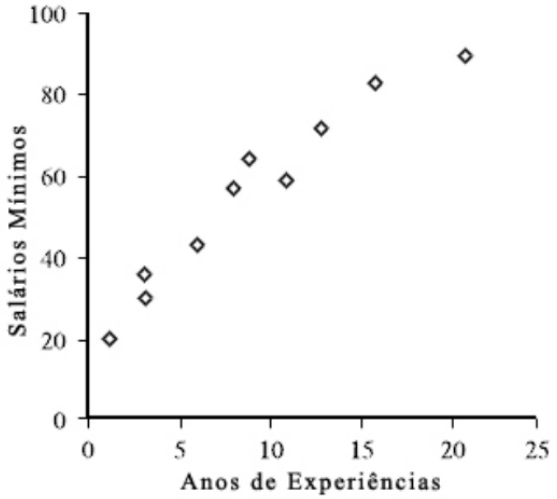

Figura 7 – Problema de Regressão Linear

Fonte: Adaptado de Han e Kamber (2005)

tende-se a pensar em bovinos, mas o termo é geral para a criação de animais.

Existem três sistemas de criação, chamados de sistema intensivo, sistema extensivo e sistema semi-intensivo (ou semi-extensivo) (ARAÚJO, 2007):

- **Sistema intensivo:** refere-se à criação do animal em confinamento, tendo normalmente mais tecnologia envolvida, mais operários e maior investimento, porém uma maior produtividade;
- **Sistema extensivo:** refere-se à criação dos animais soltos, alimentação baseada em pastagens, menor produtividade e normalmente menor tecnologia envolvida;
- **Sistema semi-intensivo:** tenta agrupar os pontos fortes dos dois outros sistemas, os animais passam a maior parte do tempo soltos, mas o investimento em tecnologia e alimentação balanceada é um fator importante.

O manejo do rebanho é muito importante para a lucratividade dos produtores, visto que o lucro está em cada animal em particular, o bom manejo torna-se essencial. O bom manejo está basicamente na organização, experiência e conhecimento técnico, para assim organizar os gastos, saber com o que gastar e conseguir tratar os animais da forma mais correta possível. O mau manejo em geral resulta apenas em pontos negativos para o produtor, pontos estes que geram a perda de produtividade e consequentemente a perda de lucro.

Na bovinocultura de leite, existem alguns dados importantes que devem ser ressaltados:

- **Período de lactação:** é o período que a vaca produz leite, gira em torno de 9 a 10 meses nas raças mais comuns de leite (holandesa e girolando), e existe o período não produtivo

(seco), que varia de raça para raça; Produção diária e total do leite: produção diária da vaca ou anual durante o período de lactação;

- **Produção diária e total do leite:** produção diária da vaca ou anual durante o período de lactação;
- **Conversão alimentar:** serve para regular o quanto o animal consome de alimento comparando com a sua produtividade, este dado torna-se importante nos casos de criação em confinamento, onde o investimento em alimentação é alto;
- **Teor de gordura:** mede-se o teor de gordura encontrado no leite produzido, fator importante para analisar a nutrição que o animal está tendo. A porcentagem varia conforme a raça, girando em torno de 4% a 6% nas raças mais comumente empregadas na produção de leite;
- **Vida útil de matrizes e reprodutores:** refere-se ao descarte de vacas improdutivas e reprodutores que não servem mais para monta ou na produção de sêmen, pois não produzem quantidade suficiente, obtiveram algum problema físico ou genético. O descarte é um fator importante, pois mantém o rebanho produtivo, garantindo uma melhor linhagem. O descarte é de cerca de 20% ao ano (ARAÚJO, 2007).

No meio pecuário são muitos os problemas que são gerados no dia a dia que afetam a produtividade das vacas. Mas o maior problema encontrado é a mastite, que é uma inflamação nas glândulas mamárias. O problema que é controlado através da Contagem de Células Somáticas (CCS), visto que o principal motivo do aumento no número de células somática é a infecção intramamária. Portanto a CCS é controlada para prever o desenvolvimento de mastite nos animais, o que afeta diretamente na produtividade, pois o tratamento contra a mastite faz com que o leite produzido seja descartado, devido aos fortes remédios aplicados nos animais (COLDEBELLA et al., 2004).

2.5 TRABALHOS CORRELATOS

Segundo Perissinotto e Moura (2007), uma equipe de Engenheiros Agrônomos da UNICAMP, realizou uma pesquisa com o objetivo de identificar o Índice Térmico de Umidade (ITU), que serve para medir a condição de estress das vacas. A pesquisa foi realizada em uma fazenda de produção leiteira comercial, sendo localizada no município de São Pedro, interior de São Paulo. Foram analisadas 15 vacas da raça holandesa, que eram tratadas em confinamento.

Para a aplicação das técnicas de mineração de dados foi utilizado o software WEKA, versão 3.4, usando o método de classificação, utilizaram o algoritmo J48 para fazer um cruzamento de dados, podendo assim gerar a árvore de decisão. A validação cruzada fez uma divisão nos dados criando 10 grupos de tamanhos iguais, sendo eles 9 de treinamento e 1 de teste. A classificação foi repetida 10 vezes usando o grupo de teste para validar o grupo de treinamento, portando as regras que foram criadas ficaram otimizadas.

Com o resultado da pesquisa, constataram que no Brasil as vacas leiteiras, por grande parte do ano, várias horas do dia, ficam sob estresse térmico, pois as médias de temperatura são sempre registradas acima da ideal. Concluíram com o trabalho realizado, que a temperatura retal e a frequência respiratória tem associação com o ITU, para vacas holandesas que ficam em confinamento.

A mineração de dados foi aprovada como uma boa ferramenta para o produtor, sendo que com as técnicas a determinação de estratégias e a tomada de decisão se tornam mais apropriadas. A ferramenta permitiu criar alguns parâmetros ideais de conforto e confinamento, para vacas em lactação da raça holandesa. Os resultados obtidos podem ser usados de várias formas, com o objetivo de melhorar as condições de ambiente do local. Exemplo, com os dados obtidos e a ajuda de um software, realizar de forma eficiente a tomada de decisão para acionar o sistema de climatização, de acordo com as condições climáticas de exposição, refletindo nos custos de produção (PERISSINOTTO; MOURA, 2007).

Fazendo a reutilização do sua primeira pesquisa como base, citada acima, o professor Maurício Perissinotto, abrangeu sua pesquisa fazendo o uso de dados de dois locais distintos, utilizando a classificação como mineração de dados e o método estatístico, lógica fuzzy, para transformar os valor quantitativos em valores simbólicos. A nova pesquisa foi realizada em uma fazenda no município de Évora, em Portugal. As condições climáticas são distintas nos dois ambientes, sendo o primeiro ambiente um clima subtropical (São Pedro) e o segundo ambiente um clima mediterrâneo (Évora).

Utilizando o conhecimento adquirido com as técnicas de classificação, geradas pelo uso do algoritmo J48 na ferramenta WEKA e cruzando dados para a geração da árvore de decisão, foi possível criar uma modelagem do sistema especialista, baseado em lógica fuzzy. Para a manipulação dos dados, foi utilizado o software MATLAB 6.1 no módulo *Fuzzy Logic Toolbox*.

Concluiu-se que, a combinação do método estatístico lógica fuzzy e da técnica de classificação de dados levou a resultados significativos para a análise de conforto térmico dos animais. Com o software MATLAB 6.1, foram estabelecidos bons parâmetros para o conforto ambiental dos animais e para o estudo realizado. Portanto, as ferramentas utilizadas e os

resultados obtidos favorecem de modo direto as formas de manejo e os custos de produção (PERISSINOTTO et al., 2009).

Segundo Naas et al. (2008), uma pesquisa realizada no ano de 2008, abrangendo o estro de bovinos leiteiros, tentou identificar técnicas de mineração de dados relevantes para a obtenção de melhores resultados. A principal técnica desenvolvida para a detecção de estro, é a visual, mas mesmo para um tratador experiente é uma tarefa difícil, e ineficiente para grandes grupos de animais. A análise foi feita em 100 vacas da raça holandesa, escolhidas aleatoriamente e tratadas em ambiente de confinamento. Para obter regras de classificação adequadas, utilizou-se árvore de decisão em conjunto com a lógica fuzzy. O atributo-alvo foi o estado da vaca estar ou não no cio. Como resultado da mineração de dados, foi comprovado que a movimentação das vacas está diretamente ligada ao estro, concordando com estudos já realizados na área. Foi possível afirmar que os cálculos gerados pela técnica de mineração de dados tornou o processo mais preciso e menos trabalhoso. Assim, melhorou-se a tomada de decisão para futuras inseminações artificiais e diminuiu-se o risco com perdas.

Em um estudo realizado na Polônia, foram analisadas 158 vacas da raça Holstein-Friesian, todas em ambiente de confinamento. Para realizar a pesquisa, foram utilizadas técnicas de regressão e árvores de classificação. O software utilizado foi o Enterprise Miner, versão 7.1, com a utilização do pacote SAS. O atributo-alvo da pesquisa foi a atividade motora dos animais na pré ordenha, no período da manhã e noturno. Para gerar uma árvore de decisão, foram realizados os cálculos do Gini Index. Pôde-se observar com a pesquisa realizada, que as primeiras vacas a ocuparem as ordenhas, eram as mais agitadas e a produção delas era maior. Para as vacas mais excitadas o tempo de ordenha foi um pouco maior, mas com o tempo de lactação a reatividade das vacas diminuiu. As técnicas de mineração mostraram-se fundamentais, pois os traços de reatividade expressaram grande complexidade e com o uso da árvore de decisão, está complexidade pôde ser bem explicada (NEJA et al., 2017).

2.6 WEKA

A ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) é um software que agrupa vários algoritmos de aprendizado de máquina e pré-processamento, gera um amplo suporte para o processo experimental de dados. Foi desenvolvida em linguagem Java por pesquisadores da Universidade de Waikato, localizada na Nova Zelândia. É um software livre,

distribuído sob os termos da *General Public License* (GPL). Testado e aprovado nas plataformas desktop mais usadas mundialmente, ou seja, em Linux, Windows e Macintosh.

Depois da ferramenta ser instalada, não é necessária nenhuma configuração adicional para a execução. Para executar o WEKA existem três opções diferentes: interface gráfica, linhas de comando e pela sua *Application Programming Interface* (API).

2.6.1 Interface do usuário

A ferramenta WEKA tem uma interface simples de ser usada, contemplando uma curva de aprendizado alta. O jeito mais fácil de usar é pela interface gráfica, Explorer, conforme a Figura 8.

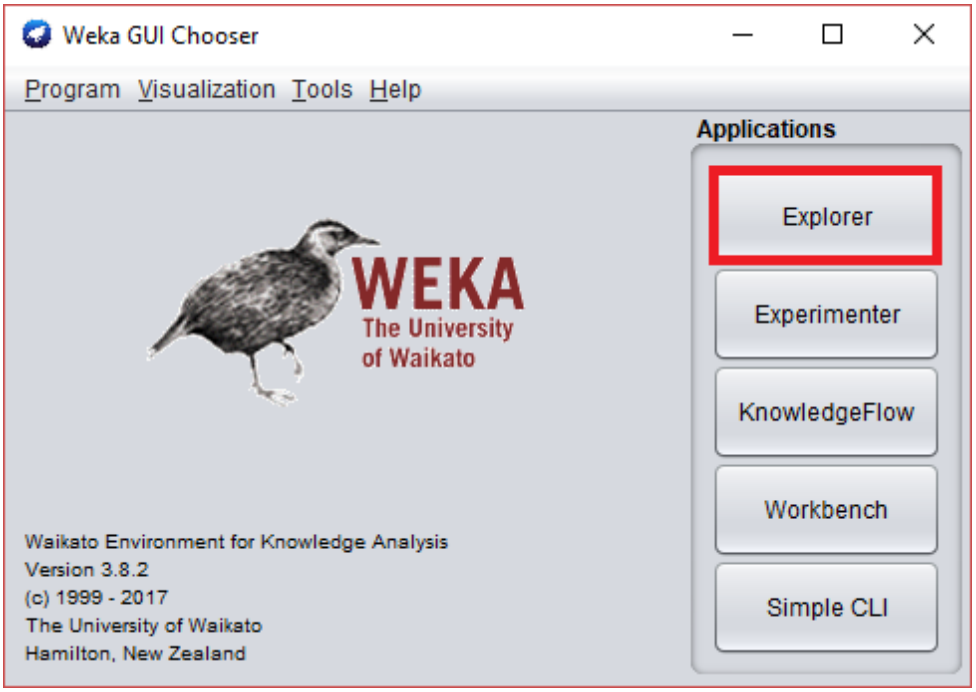

Figura 8 – Interface gráfica do WEKA, versão 3.8.2

Fonte: Autoria própria

No módulo Explorer, se obtém acesso a 6 abas distintas, como na Figura 9. Na primeira aba (Preprocess) consegue-se fazer a escolha do conjuntos de dados e modificá-lo de várias maneiras diferentes. Na segunda aba (Classify), é possível treinar os conjuntos de aprendizagem, para realizar a classificação ou regressão. Na terceira aba (Cluster), pode-se ser

capaz de aprender grupos para os conjuntos de dados. Na quarta aba (Associate), consegue-se avaliar os dados usando as regras de associação. Na quinta aba (Select attributes), pode-se selecionar os aspectos mais importantes do conjunto de dados. Na sexta e última aba (Visualize), pode-se visualizar os gráficos bidimensionais e assim conseguir interagir com eles (WITTEN et al., 2011).

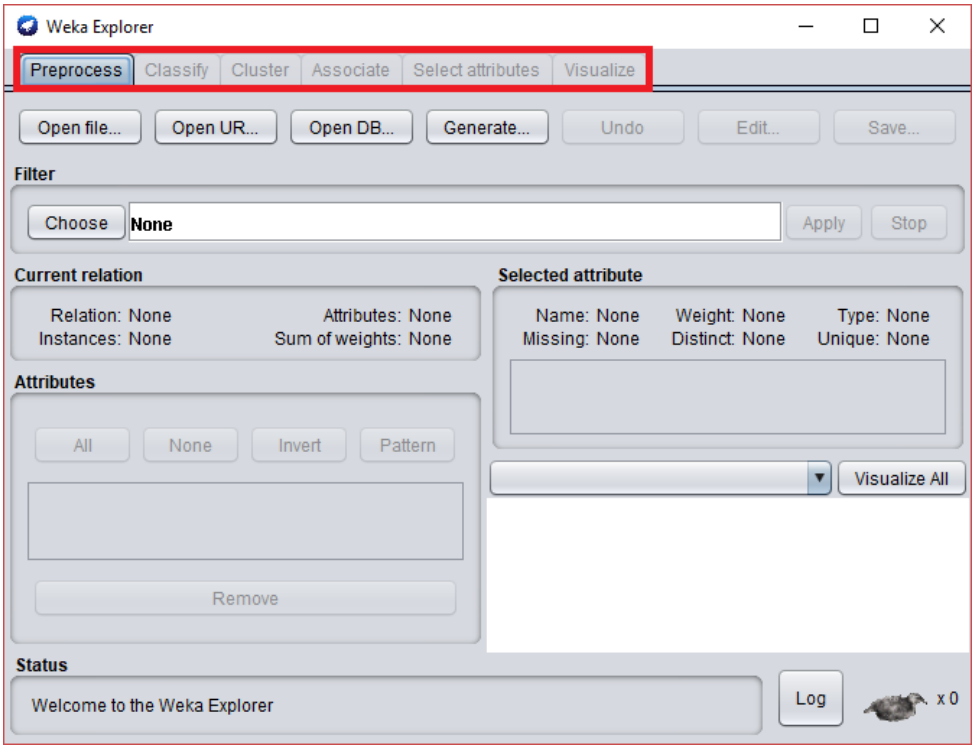

Figura 9 – Interface gráfica do WEKA Explorer

Fonte: Autoria própria

2.6.2 Algoritmo J48

O algoritmo J48 desenvolvido em Java por Quinlan em 1993, foi criado pela necessidade de uma renovação do algoritmo C4.5, que foi escrito em C. É um algoritmo que gera uma árvore de decisão através de uma base de dados fornecida, associando um conjunto de treinamento para classificar as instâncias do conjunto de testes.

Considera-se ser o algoritmo com melhores resultados na montagem da árvore de decisão, utilizando o comportamento de dividir para conquistar, onde um problema complexo

é dividido em problemas menores, usando recursivamente as estratégias para resolver os problemas menores e associando as regras aceitas por uma classe-alvo (WITTEN et al., 2011).

3 MATERIAIS E MÉTODOS

Neste capítulo, serão apresentadas as ferramentas que foram utilizadas para a análise dos dados, transformando os dados em informações válidas e de possível compreensão.

3.1 COLETA DE DADOS

A coleta de dados se mostrou muito difícil, foram várias as tentativas para encontrar produtores que obtinham a tecnologia e o controle diário de sua produção. Em alguns casos foi obtida uma resposta com vontade de ajudar, mas a base de dados não foi suficiente, em outros casos não foi enviado dados em tempo hábil para a realização do estudo. Por fim, foi encontrada uma fazenda com estrutura e interesse de ajudar na pesquisa.

A pesquisa está sendo realizada em uma fazenda de produção leiteira comercial, situada no município de Céu Azul, no interior do estado do Paraná. Fazendo uso do sistema de gerenciamento da fazenda, foi possível a coleta de dados via relatórios. Os dados coletados são dos últimos 4 anos de produção e o critério usado para a coleta de dados foi a potencial correlação dos dados com a produtividade de leite.

A coleta de dados foi feita separadamente vaca por vaca, pois o sistema de gerenciamento gera apenas o relatório geral dos últimos 10 dias de produção. Os relatórios de dados gerados estão no formato de planilha típica do software Microsoft Excel.

A fazenda conta com mais de 1100 vacas da raça holandesa, sendo que em média 500 ficam em estado produtivo durante o ano. As vacas em produção são divididas em rebanhos, que são formados conforme a Contagem de Células Somáticas (CCS). A contagem destas células se deve a uma regulamentação brasileira, onde o produtor tem que respeitar um número limite de células por mililitro de leite para poder fazer a venda do produto. A empresa conta com 8 rebanhos, que são divididos da seguinte maneira: rebanhos 1 e 2 são os pós-parto; rebanhos 3 e 4 novilhas de baixa CCS, rebanho 5 são vacas de baixa CCS, rebanhos 6 e 7 vacas e novilhas com

CCS intermediária e rebanho 8 vacas e novilhas com CCS alta. O sistema identifica cada animal por um código de quatro dígitos, começando do 0000 e continua crescente conforme os animais nascem, exceto animais comprados onde a empresa coloca um número elevado comparado ao último registrado, mantendo o controle dos animais externos.

A coleta dos dados é feita através de Identificação por Rádio Frequência (RFID), para o seu funcionamento existe um leitor de RFID, uma antena para enviar o sinal e a etiqueta com um microchip para identificar cada animal. O sistema é vigiado por um especialista, para analisar anomalias que possam identificar doenças ou perda de produtividade.

Na Tabela 1, pode-se ver o relatório dos dados diários de uma vaca, onde se tem a coluna de Dias, que conta progressivamente os dias que o sistema está implantado na empresa, a coluna Data, que registra o dia da coleta dos dados, a coluna Leite(litros), que marca a quantidade total de leite que a vaca produziu durante o dia, a coluna Atividade (passos), onde marca a movimentação da vaca em passos durante o dia, a coluna CCS que marca a quantidade de células somáticas (quantidade só é registrada quando o valor ultrapassa um milhão/mililitro), e a coluna Rest Time (min), que marca a quantidade de tempo que a vaca ficou parada em minutos durante o dia. Os dois últimos dados são coletados através de um pedômetro (equipamento que mede passos), que é colocado assim que a vaca entra em seu período de lactação.

Relatório Diário

Dias	Data	Leite (litros)	Atividade (passos)	CCS (t.l.)	Rest Time (min)
1461	3/11/2014	27,9	150		573
1462	3/12/2014	30	149		573
1463	3/13/2014	28,2	146		591
1464	3/14/2014	27,8	158		540
1465	3/15/2014	27,6	140		648
1466	3/16/2014	28,6	171		549
1467	3/17/2014	26,2	138		612
1468	3/18/2014	27,5	132		777
1469	3/19/2014	27,5	132		684

Tabela 1 – Tabela de dados diários de uma vaca

Fonte: Autoria própria

A qualidade do leite é analisada uma vez por mês, onde é coletada uma amostra por animal que está em produção. Sete atributos são levados em consideração nas análises, sendo eles: teor de gordura (% m/m), teor de proteína (% m/m), teor de lactose (% m/m), teor de sólidos totais (% m/m), teor de extrato seco desengordurado (% m/m), contagem células

somáticas (x mil/mL), teor de nitrogênio uréico (mg/dL), na Tabela 2, pode-se observar como é feita a coleta.

Relatório de Ensaio

Codigo da OS:	**OS-377949**			
Data Coleta	Data Recebimento	Data Análise	**Análises:**	
05/04/2018	**08/04/2018**	**10/04/2018**	GOR	Teor de Gordura (% m/m)
			PROT	Teor de Proteína (% m/m)
Temperatura Recebimento	**3.1 °C**		LACT	Teor de Lactose (% m/m)
Amostras Recebidas	**543**		ST	Teor de Sólidos Totais (% m/m)
Amostras não Coletadas	**1**		ESD	Teor de Extrato Seco Desengordurado (% m/m)
Amostras não Analisadas	**0**		CCS	Contagem Células Somáticas (x mil/mL)
			NU	Teor de Nitrogênio Uréico (mg/dL)

Tipo da Unidade de Análise: Animal

Amostra	Código	Identificação	GOR	PROT	LACT	ST	ESD	CCS	NU
1	3549	3549	3,7	2,91	4,6	12,22	8,52	21	12,8
2	3622	3622	3,83	3,02	4,63	12,48	8,65	40	11,2
3	3612	3612	3,74	3,46	4,41	12,61	8,87	63	11,2

Tabela 2 – Relatório mensal da qualidade do leite

Fonte: Autoria própria

A qualidade do leite é analisada para ver como está a alimentação do animal e para atender os pré-requisitos de venda. Em alguns casos o leite é descartado, e existem treze razões para isto, sendo elas: amostra sem conservante, amostra mal homogeneizada, amostra coagulada, temperatura da amostra inadequada, presença de sangue ou pus na amostra, presença de sujidades na amostra, volume insuficiente de leite, perda de amostra no transporte, não disponível - desvio operacional no laboratório, amostra congelada, amostra com prazo de validade vencido, amostra não coletada, impróprio.

Visto que na produção pecuária o clima é um fator decisivo, buscou-se dados relacionados as condições climáticas do município de Céu Azul. A primeira tentativa deu-se com o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), onde o mesmo informou que outro órgão era responsável pela coleta, monitoramento, armazenamento e distribuição dos dados, no caso o Instituto Nacional de Meteorologia (INMET), o mesmo informou não ter tais dados e indicou o Sistema Meteorológico do Paraná (SIMEPAR). Quando entrado em contato com a SIMEPAR, foi necessário um documento registrado e autenticado pela universidade, para que eles pudessem dispor dos dados e que estes dados fossem usados exclusivamente para uso acadêmico. Mas os únicos dados que foram enviados são de Cascavel - PR, que fica em média a 40KM da fazenda que está sendo realizada a pesquisa. Em uma conversa com o professor de climatologia da Universidade Tecnológica Federal do Paraná, professor Dr. Dalesio Ostrovski, o mesmo informou que os dados não podem ser considerados, visto que a distância é uma quantidade significativa, e os dados que foram coletados que são, temperatura e umidade, são

inconstantes de cidade para cidade.

Foi feita outra tentativa com uma empresa privada, no caso o Clima Tempo, onde se teve conhecimento de que a empresa tinha posse dos dados, mas o valor pedido foi muito elevado. Em uma primeira conversa o valor cobrado foi de doze mil reais, com isso foi-se tentando uma negociação, onde foi informado que os dados seriam usados exclusivamente para uso acadêmico e o valor foi reduzido para um mínimo de dois mil reais, que ainda é um valor elevado. O professor orientador também tentou entrar em contato, mas o valor final não foi alterado.

3.2 PRÉ PROCESSAMENTO

Após o término da coleta de dados, os dados coletados tem que passar por um tratamento, pois para fazer a mineração de dados na ferramenta WEKA, as informações devem estar em um formato pré-definido e algumas condições devem ser respeitadas. Como já informado os dados foram coletados através de relatórios que o sistema gera, mas os mesmos não geram todas as informações necessárias para a mineração. O dados que compunham o relatório são os dias que a vaca tem, a data que foi coletado o dado, os litros de leite que foram obtidos, os passos que a vaca andou durante o dia e os minutos em que ela ficou parada. Portanto alguns dados tiveram que ser inseridos manualmente, dados que foram coletados na própria empresa, mas que não geram automaticamente. Os dados inseridos foram: o código da vaca, a lactação e o rebanho.

Para facilitar o pré-processamento, a coleta de dados foi organizada da seguinte forma, cada rebanho foi organizado em um arquivo de excel e dentro deste cada planilha continha os dados de uma determinada vaca, as planilhas foram nomeadas com o código de cada respectivo animal.

As planilhas foram coletadas com muitos dados em branco, pois o sistema registra todos os dados desde o dia do nascimento, sendo assim foi retirado manualmente todos estes dados. Após esta limpeza, foi inserido os dados de “código” e “lactação” das vacas, uma por uma, pois cada relatório tem suas distinções. Posteriormente aos dados inseridos, usando a aba desenvolvedor que está contida no excel, pôde-se usar um macro que é gerado através da linguagem Visual Basic, que tem a função de unir todas as planilhas em uma única. Com as planilhas todas unidas conforme seus rebanhos, foi então que se teve a inserção do dado

"rebanho". Em seguida foi unificado todas os arquivos, antes separados por rebanhos, em um único arquivo com todos os dados. Esta unificação pode mudar alguns dados de lugar, portanto é necessário a verificação do documento gerado.

Como a pesquisa tem como prioridade buscar padrões na produtividade, foi criado um novo atributo para a realização da mineração. Este atributo classificou a quantidade de litros de leite, conforme as especificações que foram repassadas pelos profissionais da empresa em questão. As especificações foram: Para vacas que produziram menos de 20 litros ao dia, classificar como "Baixa", as vacas que produziram entre 20 e 36 litros, classificar como "Média" e as vacas que produziram acima de 36, classificar como "Alta". Depois de inserir o novo atributo, foi mantido um arquivo geral e um arquivo para cada rebanho, lembrando que existem 8 rebanhos, mas apenas 6 deles são produtivos.

3.3 ALGORITMOS

Foram realizados testes em condições identicas com a base de dados criada entre os algoritmos dispostos na ferramenta Weka, onde após análises de desempenho e resultados gerados, o algoritmo escolhido para encontrar os melhores classificadores foi o J48. Entre os algoritmos testados estão os algoritmos bayesianos (*BayesNet* e *NeiveBayes*) e os algoritmos de árvores de decisão (J48, LMT, *DecisionStum*p, *RandomForest*). A diferença entre eles foi pequena, mas o algoritmo J48 em todos os casos mostrou um tempo de resposta mais rápido e porcentagens de acertos maiores que os outros, portanto foi escolhido para ser utilizado na pesquisa.

3.4 AMBIENTE

Para a realização deste trabalho, está sendo utilizado um notebook com o processador Intel Core i7-4710MQ CPU 2.5GHz, 8GB de memória RAM. Neste notebook tem sido usado o sistema operacional Windows 10 64 bits, juntamente com os softwares Google Chrome, planilhas do Google Drive, Microsoft Excel e a ferramenta de mineração de dados Weka, versão

3.8.2.

4 RESULTADOS E DISCUSSÃO

A primeira mineração objetivou a obtenção de um conjunto de regras de classificação a partir da base de dados que corresponde aos dados coletados de todos os rebanhos. Nesta base, existem 296.906 registros, com uma média produtiva de 35,95 litros de leite, onde 35,62% são vacas de alta produtividade, 56,71% são de média produtividade e 7,65% são de baixa produtividade. Cada um com valores correspondentes aos seguintes atributos numéricos: Dias, Atividade, CCS, Rest, Lactação e Produtividade, sendo este último o atributo-alvo da classificação. Nesta mineração, o atributo Rebanho foi retirado, pela definição prévia de que também seriam realizadas minerações específicas para cada rebanho.

Como a ideia da pesquisa foi executar a classificação de dados a fim de identificar a possível influência dos atributos coletados no atributo-alvo Produtividade, foram utilizadas diversas configurações para a realização da mineração. Inicialmente, foi utilizado o algoritmo J48 (referente a árvore de decisão) com diferentes configurações mudando a porcentagem de treinamento e números de instâncias, para analisar a melhor abordagem em relação à base de dados considerada. Depois da realização dos testes, decidiu-se incluir neste trabalho resultados referentes a 5 valores do parâmetro "minNumObj" (número mínimo de instâncias por folha da árvore) do algoritmo J48: 4000, 5000, 6000, 7000 e 8000. Estes valores foram decididos levando em consideração o número de regras geradas após os testes de valores, com o objetivo de obter entre 10 e 20 regras de classificação válidas, sabendo que quanto maior o número de instâncias menor tende a ser o número de regras geradas. Esta quantidade de regras foi definida após conversas com os especialistas da empresa e análise dos resultados produzidos pelo algoritmo de classificação também para outros valores do parâmetro minNumObj. Ainda nas configurações do classificador J48, testou-se diferentes valores para o parâmetro "Percentage Split", que informa ao classificador a porcentagem do total de registros do conjunto de dados que ele deve usar para treinamento do classificador. Nos testes, foram usados 66%, 75% e 85%, após testar e analisar os resultados produzidos por diversos outros valores desse parâmetro.

Os resultados obtidos com os dados de todos os rebanhos, apresentados na Tabela 3, mostram que com o "Treinamento" em 75% os resultados foram melhores que os resultados obtidos com os outros valores. Mas destaca-se também que essa diferença é muito pequena.

O que vale ressaltar são as informações que encontram-se na coluna “Acerto” onde pode-se analisar quais minerações conseguiram gerar regras de classificação que proporcionaram classificações corretas para os 3 valores possíveis do atributo-alvo Produtividade (Baixa, Média e Alta). Neste sentido, as linhas marcadas como “Insatisfatório” indicam que as regras obtidas não proporcionaram esse resultado. Portanto, apenas 5 minerações alcançaram esse resultado desejado, mesmo com diferentes valores do parâmetro “Treinamento”.

Após a mineração feita na base de dados que foi coletada, o problema de não detectar a classe de baixa produtividade foi encontrado em vários classificadores. Ao analisar os dados e as porcentagens de cada classe, nota-se a disparidade entre a classe de baixa produtividade com as classes de média e alta produtividade. Para resolver este problema foi usado uma configuração da ferramenta Weka, onde ajusta-se as classes de acordo com uma porcentagem estabelecida pelo pesquisador, para que assim as classes consigam uma paridade melhor, afim de obter resultados mais justos.

Esta configuração do Weka esta localizada no pré-processamento, é um filtro supervisionado que controla as instâncias da base de dados. O filtro utilizado se chama Resample e foi configurado de forma que todas as classes ficassem uniformes, nivelando com base na classes de baixa produtividade.

Tabela 3 – Resultados para todos os rebanhos, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	4000	22	60,31%	Satisfatório	5
66%	5000	19	60,34%	Insatisfatório	5
66%	6000	19	60,15%	Insatisfatório	5
66%	7000	19	60,14%	Insatisfatório	5
66%	8000	18	60,14%	Insatisfatório	5
75%	4000	22	60,32%	Satisfatório	5
75%	5000	19	60,32%	Satisfatório	5
75%	6000	19	60,07%	Insatisfatório	5
75%	7000	19	60,07%	Insatisfatório	5
75%	8000	18	60,08%	Insatisfatório	5
85%	4000	22	60,25%	Satisfatório	5
85%	5000	19	60,10%	Satisfatório	5
85%	6000	19	60,02%	Insatisfatório	5
85%	7000	19	60,01%	Insatisfatório	5
85%	8000	18	59,98%	Insatisfatório	5

Fonte: Autoria própria

Note que o número de regras é o mesmo quando o “MinNumObj” é igual, mas o tamanho do grupo de treinamento pode fazer diferença nos resultados. No exemplo citado,

consegue-se ver que quando a mineração é feita com 66% de Treinamento e o MinNumObj é igual a 5000, obtêm-se resultados diferentes de quando o conjunto de treinamento é de 75% ou 85%.

Na sequência, empregou-se o recurso “Select Attributes” da ferramenta Weka para fazer uma seleção de atributos a serem usados na mineração de regras de classificação. As configurações usadas para essa seleção foram o método “InfoGainAttributeEval”, que usa equações numéricas para medir o ganho de informação dos atributos em relação à classe, combinado com o método “Ranker”, que faz com que os resultados fiquem organizados na forma de ranking. A junção dessas duas opções resultou na formação de um ranking dos atributos de predição que mais influenciavam nos valores do atributo-alvo. Assim, para a base de dados referente a todos os rebanhos, os atributos que se destacaram como mais relevantes foram “Dias” e “Rest”. Após obter este resultado, repetiu-se a mineração sobre esse conjunto de dados, mas utilizando como atributos de predição apenas “Dias” e “Rest”. O objetivo foi analisar se os resultados obtidos seriam similares àqueles obtidos quando foram usados todos os atributos de predição. Essa mineração foi realizada com as mesmas configurações utilizadas antes para a situação que incluiu todos os atributos. Em seguida, executou-se mais uma mineração, usando então apenas o atributo de predição considerado mais relevante (Dias).

Na Tabela 4 consegue-se ver os resultados gerados com 1 e 2 atributos, sendo eles os mais relevantes. Nota-se que os resultados foram piores que os obtidos com 5 atributos, além do fato de que todos os classificadores encontrados foram insatisfatórios, pois não conseguiram classificar vacas com baixa produtividade.

Após a obtenção dos resultados correspondentes a todos os rebanhos, foram minerados os conjuntos de dados correspondentes aos 6 rebanhos produtivos, separadamente. A base de dados do rebanho 3 conta com 58.362 registros, com uma média de produtividade de 36,10 litros de leite, onde 59,99% das vacas possuem alta produtividade, 36,22% possuem média produtividade e 6,76% possuem baixa produtividade. O que tornou necessário usar valores para o parâmetro “minNumObj” diferentes dos usados para o conjunto correspondente a todos os rebanhos. Tendo como princípio ter um número de regras entre 10 e 20, o “minNumObj” teve os seguintes valores: 500, 600, 700, 800 e 900. No entanto, foram considerados os mesmos valores anteriormente empregados para o parâmetro “Treinamento”: 66%, 75% e 85%. Os resultados dessas minerações para o rebanho 3 são mostrados na Tabela 5.

Temos o resultado com todos os atributos disponíveis, neste caso pode-se observar que, a diferença entre as porcentagens de classificações corretas também foi pequena. Mas, desta vez, foi encontrado um maior número de classificadores considerados satisfatórios (com regras prevendo os 3 valores possíveis do atributo Produtividade). Em relação ao valor de treinamento

Tabela 4 – Resultados para todos os rebanhos, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	4000	13	59,18%	Insatisfatório	2
66%	5000	10	58,73%	Insatisfatório	2
66%	6000	12	58,64%	Insatisfatório	2
66%	7000	8	58,62%	Insatisfatório	2
66%	8000	7	58,54%	Insatisfatório	2
75%	4000	13	58,98%	Insatisfatório	2
75%	5000	10	58,68%	Insatisfatório	2
75%	6000	12	58,70%	Insatisfatório	2
75%	7000	8	58,40%	Insatisfatório	2
75%	8000	7	58,47%	Insatisfatório	2
85%	4000	13	58,70%	Insatisfatório	2
85%	5000	10	58,68%	Insatisfatório	2
85%	6000	12	58,51%	Insatisfatório	2
85%	7000	8	58,36%	Insatisfatório	2
85%	8000	7	58,32%	Insatisfatório	2
66%	4000	6	57,74%	Insatisfatório	1
66%	5000	6	57,74%	Insatisfatório	1
66%	6000	6	57,74%	Insatisfatório	1
66%	7000	6	57,75%	Insatisfatório	1
66%	8000	6	57,58%	Insatisfatório	1
75%	4000	6	57,75%	Insatisfatório	1
75%	5000	6	57,75%	Insatisfatório	1
75%	6000	6	57,75%	Insatisfatório	1
75%	7000	6	57,75%	Insatisfatório	1
75%	8000	6	57,75%	Insatisfatório	1
85%	4000	6	57,51%	Insatisfatório	1
85%	5000	6	57,53%	Insatisfatório	1
85%	6000	6	57,52%	Insatisfatório	1
85%	7000	6	57,52%	Insatisfatório	1
85%	8000	6	57,52%	Insatisfatório	1

Fonte: Autoria própria

Tabela 5 – Resultados para o rebanho 3, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Tam. Arvore	Acerto	Classificador	Atributos
66%	500	36	71	62,67%	Satisfatório	5
66%	600	32	63	62,25%	Satisfatório	5
66%	700	30	59	62,07%	Satisfatório	5
66%	800	28	55	61,35%	Satisfatório	5
66%	900	24	47	61,32%	Insatisfatório	5
75%	500	36	71	62,81%	Satisfatório	5
75%	600	32	63	62,60%	Insatisfatório	5
75%	700	30	59	62,00%	Insatisfatório	5
75%	800	28	55	61,84%	Insatisfatório	5
75%	900	24	47	61,04%	Insatisfatório	5
85%	500	36	71	62,09%	Satisfatório	5
85%	600	32	63	62,55%	Insatisfatório	5
85%	700	30	59	61,93%	Insatisfatório	5
85%	800	28	55	62,26%	Insatisfatório	5
85%	900	24	47	62,15%	Insatisfatório	5

Fonte: Autoria própria

igual a 66%, apenas um classificador foi considerado inadequado. Para os valores de 75% e 85%, ocorreu o oposto, isto é, apenas um classificador foi considerado válido. Fazendo uso novamente do recurso de seleção de atributos da ferramenta Weka, obteve-se o mesmo resultado em comparação com o conjunto de dados de todos os rebanhos: os atributos Dias e Rest, nesta ordem, foram considerados os que mais exercem influência sobre o atributo-alvo. Assim, foram geradas novas minerações com esses dois atributos e apenas com o mais significativo, para analisar a possibilidade de obtenção de classificadores úteis com menos atributos.

Analisando a Tabela 6 obteve-se um maior número de classificadores satisfatórios usando 2 atributos de predição. Porém, em comparação com o uso de 5 atributos, o uso de 2 provocou uma redução no percentual de acerto desses classificadores. Os resultados com apenas 1 atributo de predição além de serem piores, foram em sua maioria insatisfatórios, contando com apenas dois classificadores que classificam todas as classes.

Para fazer a mineração do rebanho 4, onde existem 20872 registros, com uma média de 34,37 litros de leite, onde 49,65% das vacas são de alta produtividade, 43,10% são de média produtividade e 7,2% são de baixa produtividade. Foram usadas as mesmas configurações na ferramenta Weka do rebanho 3, com a diferença apenas no “MinNumObj”, pois por se tratar de uma base de dados menor o número de instâncias tende a diminuir para obtenção de melhores resultados. Seguindo o padrão adotado, primeiramente a Tabela 7 mostra os resultados com todos os atributos possíveis.

Tabela 6 – Resultados para o rebanho 3, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Tam. Arvore	Acerto	Classificador	Atributos
66%	500	19	37	59,39%	Satisfatório	2
66%	600	20	39	59,35%	Satisfatório	2
66%	700	13	25	59,27%	Insatisfatório	2
66%	800	12	23	59,25%	Insatisfatório	2
66%	900	10	19	59,25%	Insatisfatório	2
75%	500	19	37	60,04%	Satisfatório	2
75%	600	20	39	59,77%	Satisfatório	2
75%	700	13	25	59,50%	Satisfatório	2
75%	800	12	23	59,50%	Satisfatório	2
75%	900	10	19	59,40%	Satisfatório	2
85%	500	19	37	59,70%	Satisfatório	2
85%	600	20	39	59,72%	Satisfatório	2
85%	700	13	25	59,66%	Satisfatório	2
85%	800	12	23	59,43%	Satisfatório	2
85%	900	10	19	59,32%	Satisfatório	2
66%	500	17	33	59,46%	Insatisfatório	1
66%	600	12	23	59,34%	Insatisfatório	1
66%	700	11	21	59,30%	Insatisfatório	1
66%	800	9	17	59,30%	Insatisfatório	1
66%	900	9	17	59,48%	Insatisfatório	1
75%	500	17	33	59,71%	Satisfatório	1
75%	600	12	23	59,58%	Insatisfatório	1
75%	700	11	21	59,58%	Insatisfatório	1
75%	800	9	17	59,58%	Insatisfatório	1
75%	900	9	17	59,47%	Insatisfatório	1
85%	500	17	33	59,88%	Satisfatório	1
85%	600	12	23	59,80%	Insatisfatório	1
85%	700	11	21	59,98%	Insatisfatório	1
85%	800	9	17	59,65%	Insatisfatório	1
85%	900	9	17	59,80%	Insatisfatório	1

Fonte: Autoria própria

Tabela 7 – Resultados para o rebanho 4, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	50	62	65,93%	Satisfatório	5
66%	100	42	64,83%	Satisfatório	5
66%	150	30	64,67%	Satisfatório	5
66%	200	25	64,07%	Insatisfatório	5
66%	250	17	63,80%	Insatisfatório	5
75%	50	62	66,78%	Satisfatório	5
75%	100	42	65,82%	Satisfatório	5
75%	150	30	64,66%	Satisfatório	5
75%	200	25	64,18%	Insatisfatório	5
75%	250	17	64,20%	Insatisfatório	5
85%	50	62	65,82%	Satisfatório	5
85%	100	42	66,01%	Satisfatório	5
85%	150	30	65,05%	Satisfatório	5
85%	200	25	65,05%	Insatisfatório	5
85%	250	17	64,99%	Insatisfatório	5

Fonte: Autoria própria

Buscando melhores resultados e um intervalo onde boas informações fossem coletadas, o "MinNumObj" foi definido entre 50 e 250. Obtendo a marca de 66% de acerto em alguns classificadores, sendo 9 deles satisfatórios, onde o melhor classificador encontrado teve um Treinamento de 75% e MinNumObj 50. Observa-se que existe um padrão em todos os casos, onde o aumento de MinNumObj mostrou resultados piores e a partir de 200 eles se tornaram insatisfatórios. Neste rebanho mesmo com um bom resultado, foram geradas muitas regras de classificação, mostrando que existiria uma dificuldade maior em implantar estes classificadores.

Usando a seleção de atributos, os atributos Dias e Rest foram considerados os que tiveram maior influência no atributo-alvo respectivamente. A Tabela 8 exibe os resultados encontrados.

Os resultados mostram que para 2 atributos foram encontrados a mesma quantidade de classificadores adequados, e quando usado apenas 1 atributo de predição foram encontrados dois classificadores a menos. Houve uma redução na porcentagem de acerto, porém os classificadores com menos atributos se mostraram eficazes. Sendo que o mais eficaz entre eles foi o classificador com o conjunto de treinamento de 85% e 50 MinNumObj com 2 atributos de predição.

Para gerar os classificadores do rebanho 5, que conta com 33571 registros, com uma média produtiva de 35,41 litros de leite, onde 54,77% das vacas são de alta produtividade, 38,95% de média produtividade e 6,26% de baixa produtividade. Foram usadas as mesmas

Tabela 8 – Resultados para o rebanho 4, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividaden

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	50	44	63,47%	Satisfatório	2
66%	100	31	62,78%	Satisfatório	2
66%	150	25	62,78%	Satisfatório	2
66%	200	19	62,13%	Insatisfatório	2
66%	250	19	62,16%	Insatisfatório	2
75%	50	44	63,20%	Satisfatório	2
75%	100	31	62,32%	Satisfatório	2
75%	150	25	61,88%	Satisfatório	2
75%	200	19	62,16%	Insatisfatório	2
75%	250	19	62,22%	Insatisfatório	2
85%	50	44	64,26%	Satisfatório	2
85%	100	31	63,58%	Satisfatório	2
85%	150	25	63,20%	Satisfatório	2
85%	200	19	62,59%	Insatisfatório	2
85%	250	19	62,40%	Insatisfatório	2
66%	50	25	60,45%	Satisfatório	1
66%	100	20	60,23%	Satisfatório	1
66%	150	15	59,93%	Insatisfatório	1
66%	200	9	59,93%	Insatisfatório	1
66%	250	9	59,83%	Insatisfatório	1
75%	50	25	60,61%	Satisfatório	1
75%	100	20	60,54%	Satisfatório	1
75%	150	15	60,13%	Insatisfatório	1
75%	200	9	60,13%	Insatisfatório	1
75%	250	9	60,15%	Insatisfatório	1
85%	50	25	59,94%	Satisfatório	1
85%	100	20	60,14%	Satisfatório	1
85%	150	15	59,27%	Satisfatório	1
85%	200	9	59,66%	Insatisfatório	1
85%	250	9	59,66%	Insatisfatório	1

Fonte: Autoria própria

configurações na ferramenta Weka, mudando apenas o MinNumObj, que para este rebanho esteve no intervalo entre 150 e 750, mostrados na Tabela 9.

Tabela 9 – Resultados para o rebanho 5, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	150	44	63,56%	Satisfatório	5
66%	300	30	62,36%	Satisfatório	5
66%	450	22	61,08%	Satisfatório	5
66%	600	15	61,10%	Satisfatório	5
66%	750	13	61,74%	Satisfatório	5
75%	150	44	63,97%	Satisfatório	5
75%	300	30	63,10%	Satisfatório	5
75%	450	22	61,93%	Satisfatório	5
75%	600	15	62,08%	Satisfatório	5
75%	750	13	61,94%	Satisfatório	5
85%	150	44	63,82%	Satisfatório	5
85%	300	30	63,20%	Satisfatório	5
85%	450	22	61,93%	Satisfatório	5
85%	600	15	61,67%	Satisfatório	5
85%	750	13	61,67%	Satisfatório	5

Fonte: Autoria própria

Os resultados foram interessantes, visto que em nenhum dos rebanhos obteve um resultado onde todos os classificadores gerados fossem satisfatórios. O melhor resultado foi encontrado com um Treinamento de 75% e um MinNumObj de 300. Salientando que nos MinNumObj 450, 600 e 750 consegue-se entrar no padrão que o trabalho busca, que seria ter 20 regras aproximadamente com uma boa taxa de acerto.

Usando a seleção de atributos da ferramenta Weka, os mesmos resultados foram encontrados, sendo os atributos mais relevantes Dias e Rest, respectivamente. Com isso foram gerados novos classificadores, para analisar a importância deles comparada aos classificadores gerados pelos 5 atributos. A Tabela 10 demonstra os resultados obtidos.

Os classificadores gerados através do uso de 2 atributos exibem bons resultados, visto que apenas dois deles falham em gerar classificadores adequados. Nos casos onde o MinNumObj é de 750 com o Treinamento em 75% e 85%, revelam regras muito importantes, pois apenas 4 regras implicam em uma taxa de acerto de 60%, mostrando que além de serem úteis tem uma maior facilidade de aplicabilidade, afinal quanto menor o número de regras, menos condições para classificar corretamente existem. Quando usado apenas 1 atributo apenas dois dos classificadores conseguem classificar todas as classes necessárias. Mostrando a importância da combinação dos atributos na classificação. O melhor classificador encontrado

Tabela 10 – Resultados para o rebanho 5, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	150	32	62,34%	Satisfatório	2
66%	300	27	62,11%	Satisfatório	2
66%	450	13	61,25%	Satisfatório	2
66%	600	12	60,59%	Insatisfatório	2
66%	750	4	60,67%	Insatisfatório	2
75%	150	32	62,73%	Satisfatório	2
75%	300	27	62,34%	Satisfatório	2
75%	450	13	61,75%	Satisfatório	2
75%	600	12	60,78%	Satisfatório	2
75%	750	4	60,78%	Satisfatório	2
85%	150	32	62,41%	Satisfatório	2
85%	300	27	62,66%	Satisfatório	2
85%	450	13	61,75%	Satisfatório	2
85%	600	12	60,98%	Satisfatório	2
85%	750	4	60,62%	Satisfatório	2
66%	150	10	58,38%	Insatisfatório	1
66%	300	6	58,38%	Insatisfatório	1
66%	450	6	58,38%	Insatisfatório	1
66%	600	6	58,38%	Insatisfatório	1
66%	750	6	58,38%	Insatisfatório	1
75%	150	10	58,77%	Satisfatório	1
75%	300	6	58,67%	Insatisfatório	1
75%	450	6	58,50%	Insatisfatório	1
75%	600	6	58,51%	Insatisfatório	1
75%	750	6	58,51%	Insatisfatório	1
85%	150	10	58,55%	Satisfatório	1
85%	300	6	58,37%	Insatisfatório	1
85%	450	6	58,30%	Insatisfatório	1
85%	600	6	58,30%	Insatisfatório	1
85%	750	6	58,30%	Insatisfatório	1

Fonte: Autoria própria

neste caso foi o com Treinamento em 75% e MinNumObj de 150 com um acerto de 62,73%.

Para realizar a mineração do rebanho 6 onde existem 58133 registros, com uma média produtiva de 36,20 litros de leite, onde 37,61% das vacas são de alta produtividade, 54,79% são de média produtividade e 7,59% são de baixa prioridade. Foram usadas as mesmas configurações usadas até o momento, mudando apenas o MinNumObj, onde neste caso foi usado um intervalo entre 500 e 1500. Na Tabela 11 os resultados serão explanados.

Tabela 11 – Resultados para o rebanho 6, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	17	61,10%	Satisfatório	5
66%	750	13	61,74%	Satisfatório	5
66%	1000	8	61,23%	Satisfatório	5
66%	1250	5	60,73%	Insatisfatório	5
66%	1500	4	61,04%	Insatisfatório	5
75%	500	17	62,19%	Satisfatório	5
75%	750	13	61,94%	Satisfatório	5
75%	1000	8	61,51%	Satisfatório	5
75%	1250	5	61,31%	Insatisfatório	5
75%	1500	4	60,70%	Insatisfatório	5
85%	500	17	61,87%	Satisfatório	5
85%	750	13	61,67%	Satisfatório	5
85%	1000	8	61,55%	Satisfatório	5
85%	1250	5	61,13%	Satisfatório	5
85%	1500	4	60,96%	Insatisfatório	5

Fonte: Autoria própria

Os classificadores encontrados foram interessantes, visto que a porcentagem de acerto é significativa e em poucos casos foram gerados classificadores inadequados. Um ponto importante nestes resultados são as quantidades de regras que foram geradas, pois em todos os casos as quantidades são baixas, e faz com que os resultados se tornem mais atraentes. O melhor resultado obtido encontra-se quando o Treinamento é de 75% e o MinNumObj está em 500, com um acerto de 62,19%.

Usando a seleção de atributos da ferramenta Weka, foram encontrados os atributos Dias e Rest como atributos mais relevantes, respectivamente. Mantendo o padrão da pesquisa, foram gerados novos classificadores com estes atributos. A Tabela 12 exibe os resultados encontrados.

Quando usado 2 atributos foram encontrados 7 classificadores adequados, onde o melhor resultado se encontra quando o Treinamento está em 85% com um MinNumObj de 500 com uma taxa de acerto de 61,63%. Um bom resultado, visto que com 5 atributos foram gerados 10 classificadores satisfatórios, com uma taxa de acerto próxima à gerada com 2

Tabela 12 – Resultados para o rebanho 6, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	12	60,59%	Insatisfatório	2
66%	750	4	60,67%	Insatisfatório	2
66%	1000	4	60,65%	Insatisfatório	2
66%	1250	5	60,25%	Insatisfatório	2
66%	1500	4	60,47%	Insatisfatório	2
75%	500	12	61,34%	Satisfatório	2
75%	750	4	60,78%	Satisfatório	2
75%	1000	4	60,78%	Satisfatório	2
75%	1250	4	60,84%	Insatisfatório	2
75%	1500	4	60,27%	Insatisfatório	2
85%	500	12	61,63%	Satisfatório	2
85%	750	4	60,62%	Satisfatório	2
85%	1000	4	60,62%	Satisfatório	2
85%	1250	4	60,62%	Satisfatório	2
85%	1500	4	60,68%	Insatisfatório	2
66%	500	6	58,38%	Insatisfatório	1
66%	750	6	58,38%	Insatisfatório	1
66%	1000	6	58,38%	Insatisfatório	1
66%	1250	6	58,05%	Insatisfatório	1
66%	1500	2	58,05%	Insatisfatório	1
75%	500	6	58,51%	Insatisfatório	1
75%	750	6	58,51%	Insatisfatório	1
75%	1000	6	58,51%	Insatisfatório	1
75%	1250	6	58,51%	Insatisfatório	1
75%	1500	2	58,16%	Insatisfatório	1
85%	500	6	58,35%	Insatisfatório	1
85%	750	6	58,30%	Insatisfatório	1
85%	1000	6	58,30%	Insatisfatório	1
85%	1250	6	58,30%	Insatisfatório	1
85%	1500	2	58,30%	Insatisfatório	1

Fonte: Autoria própria

atributos. Em contrapartida para os classificadores gerados por 1 atributo, todos eles foram considerados inadequados, pois em nenhum caso conseguiu fazer-se a classificação de vacas com baixa produtividade.

Para a mineração do rebanho 7 que contém 65842 registros, com uma média produtiva de 37,66 litros de leite, onde 61,24% das vacas são de produtividade alta, 32,76% de produtividade média e 5,99% de produtividade baixa. Foram usadas as mesmas configurações do rebanho 6, pois o número de registros é semelhante e a busca pela quantidade de regras que é aproximadamente 20 foi alcançada. Na Tabela 13 encontram-se os resultados obtidos.

Tabela 13 – Resultados para o rebanho 7, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	32	64,96%	Insatisfatório	5
66%	750	20	64,95%	Insatisfatório	5
66%	1000	16	64,57%	Insatisfatório	5
66%	1250	12	64,10%	Insatisfatório	5
66%	1500	12	63,28%	Insatisfatório	5
75%	500	32	65,24%	Insatisfatório	5
75%	750	20	64,91%	Insatisfatório	5
75%	1000	16	64,84%	Insatisfatório	5
75%	1250	12	64,46%	Insatisfatório	5
75%	1500	12	63,92%	Insatisfatório	5
85%	500	32	64,77%	Insatisfatório	5
85%	750	20	64,55%	Insatisfatório	5
85%	1000	16	64,37%	Insatisfatório	5
85%	1250	12	64,41%	Insatisfatório	5
85%	1500	12	63,62%	Insatisfatório	5

Fonte: Autoria própria

Com facilidade observa-se que os resultados foram todos inadequados. Não conseguindo gerar nenhum classificador satisfatório, mesmo com uma taxa de acerto boa. Os classificadores só acertaram duas das três classes que existem, isto é, a classe de vacas com baixa produtividade não foi classificada. Portanto os classificadores são prejudiciais a fazenda em questão. Na Tabela 14 obtém-se o mesmo resultado, mostrando que no rebanho 7 nenhum classificador foi gerado, com todos os 5 atributos ou apenas com os atributos mais relevantes. Este resultado pode ser explicado pelo fato de que o rebanho 7 é o rebanho com a maior média produtiva. Portanto o conjunto treinamento aplicado tem poucos registros de baixa produtividade para gerar classificadores aceitáveis.

Para a mineração do rebanho 8 e último da lista, que possui 63125 registros, com uma média produtiva de 34,71 litros de leite, com uma média produtiva de 35,41 litros de

Tabela 14 – Resultados para o rebanho 7, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	17	64,11%	Insatisfatório	2
66%	750	15	63,28%	Insatisfatório	2
66%	1000	13	63,16%	Insatisfatório	2
66%	1250	13	63,04%	Insatisfatório	2
66%	1500	7	62,78%	Insatisfatório	2
75%	500	17	64,30%	Insatisfatório	2
75%	750	15	63,28%	Insatisfatório	2
75%	1000	13	63,20%	Insatisfatório	2
75%	1250	13	62,92%	Insatisfatório	2
75%	1500	7	62,70%	Insatisfatório	2
85%	500	7	64,20%	Insatisfatório	2
85%	750	17	64,07%	Insatisfatório	2
85%	1000	15	63,04%	Insatisfatório	2
85%	1250	13	62,82%	Insatisfatório	2
85%	1500	13	62,68%	Insatisfatório	2
66%	500	4	62,56%	Insatisfatório	1
66%	750	4	62,56%	Insatisfatório	1
66%	1000	4	62,56%	Insatisfatório	1
66%	1250	4	62,46%	Insatisfatório	1
66%	1500	4	61,35%	Insatisfatório	1
75%	500	4	62,87%	Insatisfatório	1
75%	750	4	62,87%	Insatisfatório	1
75%	1000	4	62,86%	Insatisfatório	1
75%	1250	4	62,85%	Insatisfatório	1
75%	1500	4	62,80%	Insatisfatório	1
85%	500	4	62,65%	Insatisfatório	1
85%	750	4	62,65%	Insatisfatório	1
85%	1000	4	62,65%	Insatisfatório	1
85%	1250	4	62,65%	Insatisfatório	1
85%	1500	4	62,55%	Insatisfatório	1

Fonte: Autoria própria

leite onde 50,65% são vacas de produtividade alta, 38,27% de produtividade média e 11,06% de produtividade baixa. Usou-se as mesmas configurações na ferramenta Weka utilizadas nos rebanhos 6 e 7, visto que o tamanho da base de dados é semelhante. Conforme observado na Tabela 15 apenas três resultados foram adequados, todos com o mesmo padrão de MinNumObj igual a 500, e com aproximadamente 60% de taxa de acerto. A quantidade de regras geradas foi um número razoável, sendo assim mostra que no rebanho 8 pode-se criar classificadores satisfatórios.

Tabela 15 – Resultados para o rebanho 8, utilizando o algoritmo J48 com 5 atributos, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	29	59,79%	Satisfatório	5
66%	750	25	58,35%	Insatisfatório	5
66%	1000	22	57,92%	Insatisfatório	5
66%	1250	14	57,63%	Insatisfatório	5
66%	1500	10	56,67%	Insatisfatório	5
75%	500	29	59,68%	Satisfatório	5
75%	750	25	58,36%	Insatisfatório	5
75%	1000	22	58,53%	Insatisfatório	5
75%	1250	14	58,28%	Insatisfatório	5
75%	1500	10	56,64%	Insatisfatório	5
85%	500	29	60,02%	Satisfatório	5
85%	750	25	59,42%	Insatisfatório	5
85%	1000	22	58,36%	Insatisfatório	5
85%	1250	14	57,72%	Insatisfatório	5
85%	1500	10	57,69%	Insatisfatório	5

Fonte: Autoria própria

Utilizando a seleção de atributos da ferramenta Weka, foi encontrado os dois atributos mais relevantes para a mineração, sendo eles os atributos Dias e o Rest, respectivamente. Sabendo destas informações, foram feitos testes para ver a relevância destes atributos em relação ao rebanho 8. Na Tabela 16 consegue-se analisar os resultados obtidos, sendo que eles foram insatisfatórios. Portanto foi impossível criar classificadores que tinham os pré-requisitos para classificar todas as classes apenas com os atributos relevantes.

Após gerar os classificadores com a base de dados original, foi feita a alteração dos dados para nivelamento das classes. Foram feitos testes apenas com os classificadores com maior taxa de acerto, eles sendo satisfatórios ou não. Foram pegos os melhores resultados de todos os rebanhos e de cada rebanho separadamente, onde apenas o melhor resultado obtido nos testes com 5 atributos, 2 atributos e 1 atributo. Após a alteração na base de dados, foi realizada a seleção de atributos da ferramenta Weka, onde foram encontrados os mesmos resultados da base

Tabela 16 – Resultados para o rebanho 8, utilizando o algoritmo J48 com 2 e 1 atributo(s) mais relevantes, tendo como atributo-alvo a produtividade

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
66%	500	18	57,10%	Insatisfatório	2
66%	750	10	57,06%	Insatisfatório	2
66%	1000	14	57,11%	Insatisfatório	2
66%	1250	4	57,23%	Insatisfatório	2
66%	1500	4	56,67%	Insatisfatório	2
75%	500	18	57,17%	Insatisfatório	2
75%	750	10	57,01%	Insatisfatório	2
75%	1000	14	57,18%	Insatisfatório	2
75%	1250	4	57,24%	Insatisfatório	2
75%	1500	4	56,67%	Insatisfatório	2
85%	500	18	56,88%	Insatisfatório	2
85%	750	10	56,85%	Insatisfatório	2
85%	1000	14	56,71%	Insatisfatório	2
85%	1250	4	56,73%	Insatisfatório	2
85%	1500	4	56,98%	Insatisfatório	2
66%	500	5	53,70%	Insatisfatório	1
66%	750	5	53,70%	Insatisfatório	1
66%	1000	5	53,71%	Insatisfatório	1
66%	1250	5	53,62%	Insatisfatório	1
66%	1500	5	53,62%	Insatisfatório	1
75%	500	5	54,02%	Insatisfatório	1
75%	750	5	54,02%	Insatisfatório	1
75%	1000	5	54,02%	Insatisfatório	1
75%	1250	5	54,05%	Insatisfatório	1
75%	1500	5	53,93%	Insatisfatório	1
85%	500	5	53,50%	Insatisfatório	1
85%	750	5	53,50%	Insatisfatório	1
85%	1000	5	53,50%	Insatisfatório	1
85%	1250	5	53,50%	Insatisfatório	1
85%	1500	5	53,49%	Insatisfatório	1

Fonte: Autoria própria

de dados original, que seriam os atributos Dias e Rest como mais relevantes, respectivamente. Na Tabela 17 pode-se os melhores resultados obtidos com todos os rebanhos na base de dados original e os dados obtidos através da base alterada.

Nota-se que os melhores resultados obtidos foram todos insatisfatórios, e na base de dados alterada o resultado foi o inverso. Houve uma perda na porcentagem de acerto, porém desta vez todas as classes foram classificadas de forma correta. Na Tabela 9b que mostra a matriz de confusão gerada, nota-se a diferença de acerto para cada classe com o classificador, confirmando que com a a base de dados modificada os resultados são mais consistentes.

Tabela 17 – Comparação dos resultados obtidos com todos os rebanhos usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
		Base de dados original			
66%	5000	19	60,34%	Insatisfatório	5
66%	4000	13	59,18%	Insatisfatório	2
75%	8000	6	57,75%	Insatisfatório	1
		Base de dados balanceados			
66%	5000	16	48,20%	Satisfatório	5
66%	4000	8	48,23%	Satisfatório	2
75%	8000	5	42,61%	Satisfatório	1

Fonte: Autoria própria

Tabela 18 – Comparação da Matriz de Confusão com os Dados Originais e Modificados de todos os rebanhos, fazendo o uso de 5 atributos.

Matriz de confusão

Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
0	7068	820	3810	2158	2385	Baixa
0	47571	9448	1789	4403	2177	Média
0	22702	13339	1086	3476	3953	Alta

Fonte: Autoria própria

Como feito anteriormente os resultados foram comparados separadamente por rebanho, começando pelo rebanho 3. Analisando a Tabela 19 apenas um dos resultados obtidos na base de dados original não era satisfatório, então a base de dados alterada mesmo com todos os classificadores sendo adequados, ainda tem uma perda considerável em taxa de acerto. Porém

o número de regras de decisão que foram criadas para a nova base de dados teve uma diminuição significativa, mostrando regras mais fortes que as antes obtidas. Contudo quando analisado a Tabela 20, onde faz-se a comparação da matriz de confusão, obteve-se um resultado mais parcial.

Tabela 19 – Comparação dos resultados obtidos com o rebanho 3 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
66%	500	71	62,67%	Satisfatório	5
75%	500	37	60,04%	Satisfatório	2
85%	700	11	59,98%	Insatisfatório	1
Base de dados balanceados					
66%	500	13	48,93%	Satisfatório	5
75%	500	14	47,73%	Satisfatório	2
85%	700	9	43,60%	Satisfatório	1

Fonte: Autoria própria

Tabela 20 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 3, fazendo o uso de 5 atributos.

Matriz de confusão

Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
190	584	582	1011	304	289	Baixa
194	2502	4504	488	542	598	Média
63	1479	9745	340	514	874	Alta

Fonte: Autoria própria

Para o rebanho 4 pode-se analisar os resultados na Tabela 21. Foram os melhores resultados obtidos pela nova base de dados, comparado os outros rebanhos, onde um classificador atingiu 62% de taxa de acerto. A mudança nas regras geradas foi pequena, porém houve uma diminuição das regras por parte dos classificadores gerados com a nova base de dados. A Tabela 22 reforça os resultados alcançados pelos dados modificados, onde demonstra classificadores mais correspondentes, portanto melhores.

Para o rebanho 5 onde os melhores resultados obtidos na base de dados original foram todos satisfatórios, os novos resultados obtidos foram todos inferiores no quesito taxa de acerto.

Tabela 21 – Comparação dos resultados obtidos com o rebanho 4 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
75%	50	62	66,78%	Satisfatório	5
85%	50	44	64,26%	Satisfatório	2
75%	50	25	60,61%	Satisfatório	1
Base de dados balanceados					
75%	50	41	62,57%	Satisfatório	5
85%	50	31	58,87%	Satisfatório	2
75%	50	15	57,36%	Satisfatório	1

Fonte: Autoria própria

Tabela 22 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 4, fazendo o uso de 5 atributos.

Matriz de confusão						
Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
64	97	207	290	85	51	Baixa
79	562	1597	87	172	177	Média
1	145	2466	31	84	327	Alta

Fonte: Autoria própria

Em contrapartida neste rebanho as regras de classificação tiveram uma diminuição dignificativa, mostrando regras mais fortes, mesmo com uma perda na taxa de acerto, observe na Tabela 23. Porém ao analisar a Tabela 24 onde faz-se a comparação da matriz de confusão, nota-se classificadores com uma proporcionalidade melhor.

Tabela 23 – Comparação dos resultados obtidos com o rebanho 5 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
75%	150	44	63,97%	Satisfatório	5
85%	150	32	62,41%	Satisfatório	2
75%	150	10	58,77%	Satisfatório	1
Base de dados balanceados					
75%	150	16	57,38%	Satisfatório	5
85%	150	15	55,91%	Satisfatório	2
75%	150	14	50,04%	Satisfatório	1

Fonte: Autoria própria

Tabela 24 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 5, fazendo o uso de 5 atributos.

Matriz de confusão

Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
110	226	173	450	129	117	Baixa
39	1261	1918	124	339	242	Média
11	657	3998	70	212	415	Alta

Fonte: Autoria própria

Em relação ao rebanho 6 analisando a Tabela 25 pode-se observar os resultados obtidos. Para os resultados com 5 e 2 atributos, com a base de dados original os resultados são mais interessantes do que com a base de dados nivelada. Porém analisando a Tabela 26 observa-se classificadores mais justos, devido a maior proporcionalidade gerada. Portanto os classificadores gerados através dos dados modificados são mais eficientes.

No caso especial do rebanho 7 onde todos os classificadores encontrados foram insatisfatórios, teve-se o resultado inverso quando usado uma base de dados modificada. Porém com uma perda de porcentagem de acerto, mas com um menor número de regras de

Tabela 25 – Comparação dos resultados obtidos com o rebanho 6 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
75%	500	17	62,19%	Satisfatório	5
85%	500	12	61,63%	Satisfatório	2
75%	1250	6	58,51%	Insatisfatório	1
Base de dados balanceados					
75%	500	12	50,83%	Satisfatório	5
85%	500	10	51,60%	Satisfatório	2
75%	1250	7	42,16%	Satisfatório	1

Fonte: Autoria própria

Tabela 26 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 6, fazendo o uso de 5 atributos.

Matriz de confusão

Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
130	157	222	418	41	237	Baixa
74	1021	2123	201	174	330	Média
31	566	4069	100	104	493	Alta

Fonte: Autoria própria

classificação, observa-se na Tabela 27. Mostrando que a paridade dos dados é importante, assim como mostrado na Tabela 28 exibindo a matriz de confusão gerada nas duas bases de dados, evidenciando uma maior proporcionalidade nos resultados, tornando os classificadores mais aptos.

Tabela 27 – Comparação dos resultados obtidos com o rebanho 7 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
66%	500	32	64,96%	Insatisfatório	5
75%	500	17	64,30%	Insatisfatório	2
75%	750	4	62,87%	Insatisfatório	1
Base de dados balanceados					
66%	500	15	46,39%	Satisfatório	5
75%	500	13	44,81%	Satisfatório	2
75%	750	11	43,59%	Satisfatório	1

Fonte: Autoria própria

Tabela 28 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 7, fazendo o uso de 5 atributos.

Matriz de confusão

Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
0	480	851	1093	411	284	Baixa
0	2092	4906	696	534	529	Média
0	1248	11789	519	424	851	Alta

Fonte: Autoria própria

Por fim, o rebanho 8, onde foi gerado apenas um classificador adequado com a base de dados original, sendo o classificador com 5 atributos, observa-se na Tabela 29. Com a base de dados modificada, todos os classificadores foram satisfatórios e analisando a Tabela 30 onde mostra as matrizes de confusão geradas nas duas base de dados, fica claro que os classificadores gerados pela base modificada são melhores, devido a maior proporcionalidade de acerto.

Após a análise de todos os resultados gerados, pode-se notar a qualidade da empresa em questão, visto que para gerar melhores resultados foi necessário um balanceamento nos dados obtidos. Os resultados mostraram que com a base de dados original foi possível a

Tabela 29 – Comparação dos resultados obtidos com o rebanho 8 usando a base de dados original e uma base de dados nivelada

Treinamento	MinNumObj	Nº Regras	Acerto	Classificador	Atributos
Base de dados original					
85%	500	29	60,02%	Satisfatório	5
75%	1250	4	57,24%	Insatisfatório	2
75%	1250	5	54,05%	Insatisfatório	1
Base de dados balanceados					
85%	500	15	48,28%	Satisfatório	5
75%	1250	7	44,63%	Satisfatório	2
75%	1250	8	39,08%	Satisfatório	1

Fonte: Autoria própria

Tabela 30 – Comparação da Matriz de Confusão com os Dados Originais e Modificados do rebanho 8, fazendo o uso de 5 atributos.

Matriz de confusão						
Dados Originais			Dados balanceados			
Baixa	Média	Alta	Baixa	Média	Alta	
65	601	399	482	146	162	Baixa
47	1827	1738	290	217	285	Média
15	985	3792	206	135	444	Alta

Fonte: Autoria própria

obteção de classificadores, mas a maioria era limitado devido a despropocionalidade da base de dados. Contudo depois da realização do balanceamento dos dados, foi possível obter bons classificadores, mostrando a importancia da ferramenta Resample do WEKA. Entre os cinco atributos que foram analisados dois destes se mostraram mais relevantes para a produtividade de leite, sendo a quantidade de dias de uma vaca e o tempo em que elas ficam paradas, isto foi possível graças a ferramenta InfoGain em conjunto com a Ranker do WEKA. Os classificadores gerados à partir da base de dados original, mostraram resultados com uma taxa de acerto próxima a 60% e os classificadores com a base balanceada uma taxa de acerto entre 40% e 60%. A diferença entre os resultados mostrou-se na matriz de confusão, onde na base de dados original existia um défciti na classificação de produtividade baixa, sendo que na base de dados balanceada este problema não ocorreu, portanto gerando classificadores e úteis.

5 CONCLUSÃO

Com os resultados obtidos é possível indicar os melhores classificadores encontrados, para todos os rebanhos como para cada rebanho separadamente. Algumas configurações dos algoritmos no Weka mostraram-se realmente úteis, sendo elas a Resample que propocionou o balanceamento dos dados e a InfoGain+Ranker que indicou os atributos mais relevantes.

Os resultados mostraram que com os dois atributos mais relevantes a taxa de acerto é considerável, mostrando assim que para empresários que estão começando a investir em técnologia seria prudente investir na captação de informações referentes a idade dos animais assim como em um pedômetro para analisar o tempo em que as vacas estão paradas. Para os resultados encontrados com a base de dados original, onde em maior parte mostrou resultados próximos a 60% de taxa de acerto, mostram-se resultados promissores analisando a quantidade de dados distintos que a base de dados possui. Porém como observado, os classificadores estavam viciados pela falta de proporcionalidade, caso que foi resolvido com o balanceamento da base de dados usando o algoritmo Resample. Após a modificação da base de dados e a geração dos novos classificadores, constatou-se que os classificadores se tornaram satisfatórios devido ao equilíbrio na taxa de acerto das classes, mostrado na matriz de confusão gerada nos testes.

Compete salientar que o objetivo da presente pesquisa, foi encontrar padrões que exercem influência na produtividade de leite e empregar da melhor forma possível os dados coletados diariamente na empresa. Neste sentido, os resultados obtidos foram satisfatórios, visto que dois dos atributos usados mostraram uma influência mais significativa que outros.

5.1 TRABALHOS FUTUROS

Após reuniões com especialistas foi informado que alguns dos atributos exercem interferência indireta na produtividade de leite. A atividade em passos coletada diariamente,

ajuda no controle do cio de cada vaca, fazendo com que a empresa fique o menor tempo possível com uma vaca sem produção. A contagem de células somáticas, ajuda no controle das inflamações intramamárias, que são as causas da mastite, que como explanado na pesquisa, força a empresa a descartar o leite produzido. A lactação ajuda no controle de nível de produtividade, mostrando o momento em que a vaca começa a perder a produtividade pela idade, facilitando no descarte.

Com essas informações pode-se concluir que apesar dos ótimos resultados encontrados, a complexidade que as informações são gerenciadas é alta. Gerando novos assuntos para pesquisas futuras. Estes assuntos são:

- A aplicação de novos algoritmos, para a comparação de resultados;
- A conciliação de dados climáticos com a produtividade leiteira;
- A mineração de dados com o intuito de ajudar a predeterminar cio em vacas usando a atividade diária;
- A correlação com diferentes raças de vacas;
- O estudo com mais de uma empresa e se possível de diferentes estados.

REFERÊNCIAS

AGGARWAL, C. C. **Data Mining: The Textbook**. [S.l.]: Springer Publishing Company, Incorporated, 2015. ISBN 3319141414, 9783319141411.

Agência de notícias do Paraná. **Paraná se mantém como segundo maior produtor de leite do País**. 2017. Disponível em: <http://www.aen.pr.gov.br/modules/noticias/article.php?storyid=95819&tit=Parana-se-mantem-como-segundo-maior-produtor-de-leite-do-Pais>.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: . San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994.

ARAÚJO, M. **Fundamentos de agronegócios**. São Paulo: Atlas, 2007.

BATISTA, G. E. de A. P. A. Pré-processamento de dados em aprendizado de máquina supervisionado. In: . São Carlos - SP: Instituto de Ciências Matemáticas e de Computação Universidade de São Paulo, 2003.

BROWN, M. S. **Data Mining For Dummies**. Canada: John Wiley and Sons Inc, 2014.

CAMILO, C. O.; SILVA, J. C. da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Goiás: Instituto de Informática Universidade Federal de Goiás, 2009.

CARVALHO, L. A. V. de. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. [S.l.]: Érica, 2001.

COLDEBELLA, A. et al. Contagem de células somáticas e produção de leite em vacas holandesas confinadas. In: . Minas Gerais: Revista Brasileira Zootec, 2004.

FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Grupo Editora Nacional, 2011. Disponível em: <https://books.google.com.br/books?id=B7BjAQAACAAJ>.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: FAYYAD, U. M. et al. (Ed.). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. Disponível em: <http://dl.acm.org/citation.cfm?id=257938.257942>.

HAN, J.; KAMBER, M. **Data mining : concepts and techniques**. San Francisco [u.a.]: Kaufmann, 2005. Disponível em: <http://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/1558604898>.

LAROSE, D. T.; LAROSE, C. D. **Data Mining and Predictive Analytics**. 2nd. ed. [S.l.]: Wiley Publishing, 2015.

MICHALSKI, R. S.; BRATKO, I.; BRATKO, A. (Ed.). **Machine Learning and Data Mining; Methods and Applications**. New York, NY, USA: John Wiley & Sons, Inc., 1998. ISBN 0471971995.

NAAS, I. de A. et al. **Estimativa de estro em vacas leiteiras utilizando métodos quatitativos preditivos**. Campinas: SciELO, 2008.

NEJA, W. et al. **The use of data mining techniques for analysing factors affecting cow reactivity during milking**. Polonia: Journal os Central European Agriculture, 2017.

PAWET, C. **Data Mining Algorithms: Explained Using R**. Polônia: Department of Electronics and Information Technology Warsaw University of Technology, 2015.

PERISSINOTTO, M.; MOURA, D. J. de. Utilização do conforto térmico de vacas leiteiras utilizando mineração de dados. In: . Campinas: Revista Brasileira de Biossistemas, 2007.

PERISSINOTTO, M. et al. Conforto térmico de bovinos leiteiros confinados em clima subtropical e mediterrâneo pela análise de parâmetros fisiológicos utilizando a teoria dos conjuntos fuzzy. In: . Santa Maria: Scielo, 2009.

Universo Online. **Receita das exportações brasileiras de lácteos recuou 34% em 2017**. 2018. Disponível em: <https://sfagro.uol.com.br/exportacoes-brasileiras-lacteos-2017/>.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560.

YE, N. **Data Mining: Theories, Algorithms, and Examples**. 1st. ed. Boca Raton, FL, USA: CRC Press, Inc., 2013. ISBN 1439808384, 9781439808382.