UNIVERSIDADE FEDERAL DO PARÁ
CAMPUS UNIVERSITÁRIO DE CASTANHAL
FACULDADE DE COMPUTAÇÃO

SANDIO MACIEL DOS SANTOS

DESCOBERTA DE PADRÕES NO PERÍODO REPRODUTIVO DE MACACOS DA ESPÉCIE *SAIMIRI COLLINSI* ATRAVÉS DA MINERAÇÃO DE DADOS.

Castanhal – Pará
2017

SANDIO MACIEL DOS SANTOS

DESCOBERTA DE PADRÕES NO PERÍODO REPRODUTIVO DE MACACOS DA ESPÉCIE *SAIMIRI COLLINSI* ATRAVÉS DA MINERAÇÃO DE DADOS.

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação do Campus de Castanhal da Universidade Federal do Pará, como requisito para obtenção de título de Bacharel em Sistemas de Informação.

Orientador(a): Prof.ª. Drª. Fabíola Pantoja Oliveira Araújo.

Castanhal – Pará
2017

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

M152d Maciel dos Santos, Sandio.
Descoberta de padrões no período reprodutivo de macacos da espécie saimiri collinsi através da mineração de dados / Sandio Maciel dos Santos. — 2017.
55 f. : il. color.

Orientador(a): Profª. Dra. Fabíola Pantoja Oliveira Araújo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Pará, Campus Universitário de Castanhal, Faculdade de Sistemas de Informação, Castanhal, 2017.

1. Descobertas de padrões. 2. Mineração de dados. 3. Primatas Não-Humanos. 4. Medicina Veterinaria. I. Título.

CDD 006.312

UNIVERSIDADE FEDERAL DO PARÁ

SANDIO MACIEL DOS SANTOS

DESCOBERTA DE PADRÕES NO PERÍODO REPRODUTIVO DE MACACOS DA ESPÉCIE *SAIMIRI COLLINSI* ATRAVÉS DA MINERAÇÃO DE DADOS.

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação do Campus de Castanhal da Universidade Federal do Pará, como requisito para obtenção de título de Bacharel em Sistemas de Informação.

Orientador(a): Prof.ª. Drª. Fabíola Pantoja Oliveira Araújo.

Aprovado em: 16/10/2017.

BANCA EXAMINADORA

_____

Prof.ª Drª. Fabiola Pantoja Oliveira Araújo (Orientadora)

_____

Prof.ª Drª. Yomara Pinheiro Pires (Examinadora Interna)

_____

Prof. Dr. Diogo Acatauassú Nunes (Examinador Interno)

Castanhal – Pará
2017

“Mas aqueles que esperam no senhor renovam as suas forças. Voam alto como águias e não ficam exaustos, andam e não se cansam”.

(Isaías 40:31)

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus, que durante toda a minha vida, foi o Deus com um amor incalculável e, um pai de imensurável proteção, que me proporcionou inúmeras conquistas e graças. Mostrando que a paciência é dom necessário para que eu pudesse entender melhor o tempo certo para que cada sonho fosse alcançado, mesmo que eu ainda não o entendesse seu propósito e lhe agradece-se.

Gostaria também de agradece a minha mãe Sônia Maria, a qual é responsável por essa conquista, pois sempre foi um exemplo de determinação e dedicação em todas as áreas da minha vida, através de ensinamentos de conselhos que foram de suma importância para construção do caráter e valores éticos e morais. Apesar de eu nem sempre ser um filho compreensivo. Mas quero que saiba que independentemente de tudo, você é muito amada. É claro que palavras nunca serão suficientes para me expressar sobre você.

Aos irmãos: Celso, Sérgio e Suelane, eu só tenho a agradecer pelos incentivos, brigas e principalmente por cada ensinamento apreendido com vocês. Aos meus Familiares como um todo: tios, primos, avós, que sempre contribuíram de forma direta ou indireta com essa conquista.

A turma de sistemas de informações 2013.4, obrigado pela paciência, ajuda e principalmente pelo companheirismo que essa turma sempre demonstrou, nossa turma foi e vai ser a melhor, pois aqueles que tiveram contato nunca excitaram em voltar, pois sempre foram bem acolhidos por todos.

Aos professores pelos ensinamentos ao longo desses quatro anos de graduação, em especial a minha orientada Fabiola Araújo, Wlaisa Sampaio e a minha queria professora Penha Harb, que sempre me incentivaram e me cobraram bons resultados, vocês sempre serão lembrados como profissionais excepcionais e pessoas maravilhosas sempre dispostas a ajudar.

Aos meus amigos como um todo não irei citar nomes, pois seria uma nova monografia, apenas farei os agradecimentos, pois cada um contribuiu de forma significativa em minha vida pessoal, profissional e acadêmica. Espero também de alguma forma ter contribuído positivamente, torço pelo sucesso de cada um de vocês.

Por fim, reconheço todos os meus esforços, perseverança e dedicação que foi atribuída durante essa jornada de trabalhos, tanto acadêmicos, quantos profissionais, pesquisas até esse momento da minha vida, pois só eu sei quais dificuldades tiveram que ser vencidas com muito suor e orgulho nesse longo caminho. Hoje me sinto um privilegiado por Deus ter colocado todos esses obstáculos em minha trajetória e ter me dado forças para vencê-los.

RESUMO

Ainda hoje algumas das análises em bases de dados têm sido realizadas de forma manual ou quase sem uso de tecnologias especializadas, como é o caso dos pesquisadores de Medicina Veterinária da Universidade Federal do Pará – campus Castanhal, que realizam suas análises através de cálculos estatísticos por meio de planilhas, demando bastante tempo em suas análises até que as informações úteis sejam de fato detectadas. Assim o presente trabalho descreve aplicação do processo de descoberta de conhecimento em base de dados sob o conjunto de dados dos pesquisadores de Medicina Veterinária para resolver um problema real relacionado à detecção de padrões através dos características físicas e morfológicas dos macacos da espécie *Saimiri Collinsi* que vivem em cativeiro. Com ênfase na etapa de mineração de dados visando solucionar as seguintes premissas: determinar quais características físicas (dobras cutâneas do braço e tórax, peso e o volume testicular) têm maior influência na estação reprodutiva dos animais. Após é feita uma nova análise sob os parâmetros de qualidade seminal com o objetivo de identificar qual o melhor período para a coleta do sêmen dos macacos. Todo esse processo foi realizado através das regras de associação por meio dos algoritmos *Apriori* e o *FP-Growth*, para uma melhor performance dos resultados obtidos através de uma comparação algorítmica. Vale ressaltar, que houver a necessidade de gerar dados sintéticos (DS) devido à pouca quantidade de dados disponíveis para realizar o treinamento dos algoritmos. Sendo assim cada resultado obtido nessa pesquisa foi analisado e avaliado juntamente aos especialistas da área, para que não houvessem dúvidas, nem interpretações equivocadas garantindo, assim, a integridade dos resultados.

Palavras chaves: Descobertas de padrões, Mineração de dados, Primatas Não-humanos e Medicina Veterinária.

ABSTRACT

Even today, some of the basic analyzes have been performed manually or almost without using specialized technologies, such as the Veterinary Medicine researchers of the Federal University of Pará - Campus Castanhal, who carry out their analyzes through statistical calculations using spreadsheets, taking enough time to analyze them until useful information is actually detected. Thus, the present work describes application knowledge discovery in Database's process under dataset of researchers of Veterinary Medicine to solve a real problem related to the detection of patterns through the physical and morphological characteristics of monkeys of the species Saimiri Collinsi living in captivity. With emphasis on the data mining stage, the following assumptions are determined: determine which physical characteristics (arm and chest skinfolds, weight and testicular volume) have a greater influence on the reproductive season of animals. After, a new analysis under the parameters of seminal quality is made as the objective is to identify the best period for the collection of semen of the monkeys. All this process was performed through the association rules through the Apriori and the FP-Growth algorithms, for a better performance of the results obtained through an algorithmic comparison. It is worth mentioning that there is a need to generate synthetic data (DS) due to the small amount of data available to perform the algorithm training. Thus, each result obtained in this research was analyzed and evaluated together with the experts of the area, so that there were no doubts or misinterpretations thus ensuring their integrity.

Keywords: Pattern Discoveries, Data Mining, Nonhuman Primates and Veterinary Medicine.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ATP	Adenosina trifosfato
CD:	Cauda dobrada
CFE:	Cauda fortemente dobrada
CI:	Cabeça Isolada
CE:	Cauda enrolada
C_{MIN}	Confiança
DR:	Dados reais
DS:	Dados Sintéticos
EE:	EletroEjaculação
FATTED:	Período reprodutivo
FP-GROWTH:	*Frequente Partten Mining*
GAUSS:	Gaussiana
GCD:	Gota citoplasmática distal
GCP:	Gota citoplasmática proximal
IN:	Sêmen natural
-INF:	Conjunto de dados inferiores

INF+:	Conjunto de dados superiores
ITEMSET:	Itens repetidos
KDD:	*Knowledge Discovery in Database*
LK:	Lista de itens
MCC:	Microcefalia
MACP	Média Aritmética de Cada Período
MEAN:	Média
NO-FATTED:	Período não reprodutivo
NORMRND:	Função responsável por gera dados sintéticos
PG:	Pseudogota
PID:	Peça intermediaria dobrada
PIQ:	Peça intermediaria quebrada
PSEM:	Ejaculação sem propriedades bioquímicas
S_{MIN}:	Suporte
STD:	Desvio padrão
TID:	Transações
VE:	Vidro estimulação peniana
WEKA:	*Weikato Enviroment for Knowledge Analysis*

SUMÁRIO

SEÇÃO 1

INTRODUÇÃO

Nesta seção, são apresentadas a problemática do trabalho, as motivações e justificativas, para que o mesmo fosse desenvolvido, bem como os objetivos gerais e específicos do trabalho e sua estrutura de organização.

1.1. PROBLEMÁTICA

Com a disponibilidade e acessibilidade dos recursos tecnológicos, hoje nota-se que o volume de dados gerados e armazenados tem crescido expressivamente nas últimas décadas, sendo eles providos por organizações e/ou instituições públicas e privadas em geral. Assim, a extração de informação em bases de dados não automatizada torna-se inviável e complexa.

Segundo (GOLDSCHMIDT et al., 2015, 1p) o valor dos dados armazenados está tipicamente ligado à capacidade de se extrair conhecimento de mais alto nível a partir deles, ou seja, informação útil que sirva para o apoio à tomada de decisão, e/ou para exploração e melhor entendimento do fenômeno gerado de dados.

Cada vez mais, novas tecnologias têm sido desenvolvidas para solucionar a problemática apresentada referente à extração de informação em bases dados úteis para a sociedade como um todo. Com isso, as técnicas de mineração de dados têm possibilitado que diversas áreas do conhecimento apliquem suas metodologias na busca de novos conhecimentos em bases de dados.

No contexto de Medicina Veterinária, especificamente na análise de dados, há uma grande dificuldade relacionada com às técnicas e as ferramentas adequadas que automatizem e auxiliem as análises dos dados coletados através de suas pesquisas, isso tem impedido que a validação e a confirmação dessas suspeitas levantadas sejam constatadas devido a uma limitação frente ao uso de métodos não automatizados, tais como: estatística e a matemática convencional.

Mediante ao que foi apresentado acima, o presente trabalho utiliza as técnicas de mineração de dados com a finalidade de identificar padrões estatísticos relacionados aos períodos não reprodutivo e reprodutivo dos macacos da espécie *Saimiri Collinsi* (macacos-de-cheiro), que vivem em cativeiro, norteando alterações anatômicas, tais como: o aumento das dobras cutâneas do braço, tórax, peso e o volume testicular desses animais.

1.2. CONTRIBUIÇÕES DO TRABALHO

As contribuições deste trabalho são: automatização do processo, por meio da descoberta de conhecimento em bases de dados - KDD, de análise dos dados coletados sobre os macacos da espécie *Saimiri Collinsi*, que vivem em cativeiro, identificando o seu ciclo de reprodução; descoberta de padrões, até então não percebidos pelas ferramentas tradicionais de analise; redução do tempo de análise dos dados; constatação de paradigmas referentes ao período reprodutivo até vivenciados.

1.3. JUSTIFICATIVA E MOTIVAÇÃO

O processo de obtenção e acúmulo de dados realizado pelos pesquisadores dos cursos de Medicina Veterinária da Universidade Federal do Pará Campus Castanhal, tem sido realizado em planilhas eletrônicas, que pode ser considerado como um processo parcialmente automatizado. Porém, mesmo com essa semi-automatização no processo de armazenamento de dados as suas análises estatísticas não tem se mostrado tão eficiente na extração de conhecimento em bases de dados.

Com isso, os pesquisadores da área possuem, atualmente, uma grande dificuldade em encontrar padrões estatísticos baseados nas características físicas e morfológicas desses animais capazes de identificar quando um macaco da espécie *Saimiri Collinsi*, que vive em cativeiro, encontra-se em período não reprodutivo ou reprodutivo. Assim, algumas medidas anatômicas como: volume testicular, dobras cutâneas dos braços e do tórax e o peso foram utilizadas para auxiliar na detecção desses padrões que identifiquem se o animal está ou não em seu período reprodutivo.

No entanto, no decorrer deste trabalho alguns dados sintéticos tiveram que ser gerados pois a quantidade de dados disponíveis para realizar a mineração de dados era pequena, o que poderia acarretar em uma menor confiabilidade nos resultados obtidos. Outro ponto importante a ser ressaltado sobre a geração de dados artificiais está relacionado com a captura de animais na fase adulta que podem apresentar um curto tempo de vida em cativeiro, pois assim não se consegue colher uma quantidade grande de amostras devido a isso.

Sendo assim, este trabalho auxilia os pesquisadores da área na descoberta de padrões referentes aos dados coletados dos macacos da espécie *Saimiri Collinsi*, que vivem em cativeiro, ou melhor, este trabalho utiliza técnicas automatizadas de mineração de dados para detectar padrões relacionados aos períodos não reprodutivo e reprodutivos dos animais bem como, o melhor período para a coleta do sêmen.

1.4. OBJETIVOS

1.4.1. OBJETIVO GERAL

O objetivo principal deste trabalho é apresentar o uso de técnicas de mineração de dados para identificar padrões macro observáveis nos macacos do gênero *Saimiri Collinsi,* que vivem em cativeiro, pois esses não apresentam os mesmos padrões daqueles que vivem livre na natureza. Essa análise leva em consideração algumas características físicas e morfológicas que cada um pode apresentar durante o período de reprodução.

1.4.2. OBJETIVO ESPECÍFICO

- Realizar o pré-processamentos das bases de dados e validar qualquer alteração feita, como: remoção de dados duplicados e a remoção de dados em branco juntamente aos especialistas da área.
- Detectar padrões através da tarefa de associação e constatá-las mediante aos especialistas da área.
- Gerar dados artificiais a partir de estatísticas dos dados coletados.
- Testar e avaliar algoritmos de regras de associação que estão implementados na ferramenta Weka com os dados gerados.

- Desenvolver relatórios e gerar gráficos dos resultados, para que os avaliadores da área de Medicina Veterinária possam interpretá-los de maneira clara e objetiva, com isso comprovar as suposições levantadas no decorrer deste trabalho.

1.4. ESTRUTURA DO TRABALHO

Este trabalho está organizado e composto por cinco seções, incluído esta seção introdutória. Sendo as demais seções estruturadas da seguinte forma:

Seção 2: Esta seção, apresenta conceitos referentes a ecologia e o gênero *Saimiri* em vida livre, bem como aspectos e técnicas de coleta do sêmen em períodos reprodutivos dos animas e descreve brevemente sobre qualidade seminal.

Seção 3: Nesta seção, são apresentados alguns conceitos básicos sobre o processo de Descoberta de Conhecimento em Bases de Dados, bem como a tarefa de mineração de dados e suas técnicas, descrevendo os algoritmos *Apriori* e o *FP-Growth* e suas aplicações.

Seção 4: Nessa seção, são apresentadas as fases de preparação dos dados para aplicação das regras de associação em um problema real. Além de descrever cada padrão encontrado através testes realizados nos parâmetros físicos e morfológicos dos macacos-de-cheiro que vivem em cativeiro.

Seção 5: Nesta seção, são apresentadas as considerações finais e as dificuldades encontradas para realização deste trabalho. Consequentemente, são sugeridos alguns trabalhos futuros que podem ser realizados nesta área de conhecimento.

SEÇÃO 2

FUNDAMENTAÇÃO TEÓRICA SOBRE *SAIMIRI COLLINSI*

Esta seção apresenta conceitos referentes a ecologia e o gênero *Saimiris* em vida livre, bem como aspectos e técnicas de coleta do sêmen em períodos reprodutivos dos animas e descreve brevemente sobre qualidade seminal.

2.1. ECOLOGIA DO GÊNERO DE PRIMATAS NEOTROPICAIS *SAIMIRI*

Os *Saimiris* são primatas não humanos também conhecido popularmente no território brasileiro como macaco-de-cheiro ou mão-de-ouro (AURICCHIO 1995). Encontrados frequentemente na Amazônia brasileira, na Venezuela, no Peru e na Bolívia entre outros países da América do Sul (BALDWIN & BALDWIN, 1981; HERSHKOVITZ 1984; JACK, 2007 Apud PAULINHO, 2013, 8p), são animais de pequeno porte, possuem hábitos diurnos (INGBERMANN; STONE; CHEIDA, 2008), vivem em pequenos grupos com múltiplos machos e múltiplas fêmeas (GOODALL; MITTERMEIER, 1999; FORTMAN et al*.,* 2002 apud VIANA, 2013, 27p) e podem chegar aos 20 anos. Porém, estima-se que alguns possam chegar e passar dos 30 anos (WILLIMS, 2008).

Esses animais apresentam pelagem acinzentada a esverdeada com tons amarelos e laranjadas em suas patas que pode variar de acordo com a espécie estudada (GROVES, 2005 Apud VIANA, 2013, 19p), conforme a Figura 1.

Figura 1. Exemplar do gênero *Saimiri*

Fonte: Extraído do (ICMBio, 2107)

Os *Saimiris* são compostos por cinco espécies: *Saimiri sciureus, S. ustus, S. boliviensis, S. oerstedii e S. vanzolinii (*RYLANDS; MITTERMEIER, 2009 apud VIANA, 2013*),* onde todas as espécies citadas acima apresentam um declínio populacional em vida livre (IUCN, 2009 apud KUGELMEIER, 2011). Segundo os mesmos autores, as espécies *S. oerstedii e a S. vanzolinii,* são consideradas vulneráveis, já as espécies *S. sciureus, S. ustus* e *S. boliviensis* não estão ameaçadas.

2.2. ASPECTOS REPRODUTIVOS DO GÊNERO *SAIMIRI*

O processo reprodutivo de primatas não humanos é marcado por inúmeras fases, que inicia na etapa de cortejo do macho para com a fêmea, estendendo-se até o acasalamento assim outras características podem ocorrer neste ciclo como a prole (KUGELMEIER, VALLE, MONTERIO, 2010, 61p). Sendo assim, a atividade reprodutiva é de extrema importância, pois para qualquer espécie é o fator que assegura a progressão da espécie no meio (KUGELMEIER, VALLE, MONTERIO, 2010, 62p).

Dentro do período reprodutivo dos *Saimiris* o macho e fêmea apresentam poucas diferenças macro observáveis em seus aspectos físicos, no entanto, o aumento de massa corporal no período reprodutivo é um acontecimento que fica mais visível nos machos, este período também é denominado "*fatting"* (DUMOND; HUTCHINSON, 1967 apud VIANA, 2013). Esse acumulo de gordura e água que ocorre principalmente no subcutâneo do tórax, ombros e braços, conferindo-lhes uma aparência de animais gordos (BALDWIN, 1985).

Outro quesito no que diz respeito à reprodução para o gênero *Saimiri é* o sistema de acasalamento de múltiplos parceiros, onde vários machos copulam como diferentes fêmeas ocorrendo seleção sexual via competição espermática (KUGELMEIER, VALLE, MONTERIO, 2010, 62p). Neste sistema de Multimachos-Multifêmeas a relação sexual é rápida, não exclusiva e gregária outros exemplos de primatas que seguem esse mesmo padrão: *gênero Macaca, Pan, maioria do Papio* (DIXSON E ANDERSON, 2001).

A maturidade sexual no *Saimiri* ocorre de 2,5 a 3,5 anos de idade nas fêmeas (Richter, Lehner & Hendrickson, 1984, apud KUGELMEIER, VALLE, MONTERIO,2010, 68p) e nos machos é um pouco mais tardia ocorrendo a partir dos 3,5 a 5 anos de idade (DUKELOW, 1983; BALDWIN, 1985 Apud KUGELMEIER, 2011, 24p). Essa maturidade antecipada das fêmeas está correlacionada ao seu peso, outro fato relevante é o ciclo ovariano ser mais curto perto a outras espécies de primatas com o período de duração de 7 a 12 dias e outros de 8 a 10 dias (DUKELOW, 1983; BALDWIN, 1985 Apud KUGELMEIER, 2011. 24p), sua gestação dura em média 156 dias depois do acasalamento e, normalmente nasce um filhote com 15% do peso da mãe (BALDWIN, 1985).

Os machos maduros sexualmente apresentam uma bolsa escrotal semi-pendulosa e assimétrica proporcionalmente grande pelo seu porte (STEINBERG et al., 2005). O testículo e o epidídimo juntos formam uma massa relativamente pequena e globular, o seu pênis é cilíndrico e mede 35 mm de comprimento encoberto por prepúcio retrátil também apresenta osso peniano e algumas espículas queratinizadas vestigiais laterais ao seu corpo (Hill, 1960 apud STEINBERG et al., 2005).

2.2.1. SAZONALIDADE REPRODUTIVA EM *SAIMIRI*

As espécies de macacos-de-cheiro têm períodos sazonais bem definidos (COE; ROSENBLUN, 1978; BALDWIN, 1985 Apud VIANA, 2013, 24p), assim a estação de acasalamento frequentemente ocorrerem em períodos não chuvosos e os partos normalmente ocorrem em estações chuvosas, isso está correlacionado a alguns fatores ambientais como: índice pluviométrico, latitude, temperatura e o fotoperíodo (tempo que uma planta ou animal precisa ficar exposto a luz solar), podendo também ser relacionado a disponibilidade de alimentos que é mais abundante nessas épocas (NELSON et al., 1990; MALPAUX, 2006 apud VIANA, 2013, 23p). Embora esses fenômenos ambientais e climáticos possam afetar os períodos não reprodutivo e reprodutivo desses animais não há uma estimativa do real impacto que essas variáveis possam ter sobre eles (COE; ROSENBLUM, 1978 Apud KUGELMEIER, 2011).

Uma mudança morfológica marcante anteriormente mencionada é o fenômeno "*fatting*" (MENDOZA et al., 1978; MENDOZA, 1999 Apud KUGELMEIER, 2011, 26p) normalmente 2 a 3 meses antes da estação reprodutiva juntamente com o ganho de massa corporal que pode variar de 85 a 300g, esse aumento de peso varia conforme a espécie e o porte do animal (DuMOND, 1968; BALDWIN, 1985 Apud KUGELMEIER, 2011, 26p). Concomitante ao aumento de massa corpórea há também um aumento das dobras cutâneas do braço e tórax essas características chamam bastante atenção visualmente dos especialistas (BALDWIN, 1985).

Outro ponto a ser levado em consideração está relacionado aos aspectos fisiológicos juntamente com a variação de concentração de alguns hormônios que podem ter relação com o "*fatting*" e com a estação reprodutiva da produção e qualidade do sêmen desses primatas (COE; ROSENBLUM, 1978 Apud KUGELMEIER, 2011). Em relação ao fator comportamental há um aumento na interação macho-fêmea no período de acasalamento onde os machos aproximam-se das fêmeas com uma maior frequência (BALDWIN, 1985). Com isso, os machos normalmente apresentam alguns sinais externos como odor e olfato nessa época reprodutiva, na tentativa de realizar o acasalamento. (SCHIML et al., 1999 Apud VIANA, 2013, 25p).

2.3. COLHEITA E AVALIAÇÃO DO SÊMEN EM PRIMATAS NÃO HUMANOS

Em primatas não humanos a colheita pode ser feita de inúmeras formas, tais como: masturbação, lavagem do canal após a cópula, uso de vaginas artificiais, eletroejaculação por via retal (EE), vidro estimulação peniana (VE) ou punção realizada diretamente no epidídimo. Para isso, as várias técnicas de obtenção do sêmen que podem ser utilizadas, no entanto, cada uma das técnicas descritas pode ou não gerar alguns desconfortos nos animais (DUKELOW, 1971; GOULD; MARTIN, 1986; WILDT, 1986, 1996; MORREL; HODGES, 1998; MORREL; HODGES, 2001; VALLE, 2007 Apud KUGELMEIER, 2011, 36p). Assim, o método de eletroejaculação (EE) é um dos mais utilizados na área, o mesmo é responsável pela maior parte das experiências e trabalhos realizados em primatas não humanos principalmente em micos-de-cheiro (YEOMAN; 1998), mesmo que haja alguns desconfortos ocasionados pela aplicação dessa técnica segundo o mesmo autor citado anteriormente.

A avaliação do sêmen é de extrema importância, pois é através dela que a qualidade espermática é diagnosticada e essa qualidade é de suma importância para que possa ocorrer a fertilização natural ou artificial. Com isso, a avaliação espermática tenta determinar o potencial de fertilização que cada amostra seminal possui, seja ela: fresca, resfriada ou descongelada. Assim, é necessário que se utilize métodos rápidos com custo-benefício acessível (MOCÈ; GRAHAM, 2008 apud KUGELMEIER, 2011, 42p).

As etapas de avaliação do sêmen são compostas por: testes físicos, morfológicos e funcionais (VIANA, 2013, 61p), onde cada um desses métodos de análise mencionados é responsável por uma atividade específica de avaliação do sêmen. Os testes físicos ocorrem logo após a etapa de colheita que consistem na avaliação da aparência do ejaculado: pH, volume, concentração espermática, motilidade, vigor, e movimentos progressivos retilíneos (WILDT, 1986; VALLE, 2007; MOCE; GRAHAM, 2008 apud KUGELMEIER, 2011, 43p). Avaliação morfológica espermática pode ser avaliada visualmente através do uso de reagente ou simplesmente sem auxílio de qualquer ferramenta automatizada (MOCÈ; GRAHAM, 2008 apud KUGELMEIER, 2011, 43p), essa avaliação tem intuito de determinar quais espermatozoides apresentam-se anormais e quais são ditos normais (KHOLKUTE; GOPALKRISHNAN; PURI, 2000 apud KUGELMEIER, 2011, 43p). E, por fim não

menos importante, temos os testes funcionais que são responsáveis pela avaliação da integridade da membrana plasmática que é da atividade mitocondrial (KUGELMEIER, 2011,43p). Assim, a integridade da membrana plasmática é responsável por regular a homeostase atuando como agente principal na sobrevivência do espermatozoide no interior do trato reprodutor feminino e na capacidade de fertilização (ANDRADE et. al., 2007). Segundo o mesmo autor a atividade mitocondrial é responsável pelo ATP e por isso estão diretamente relacionadas ao movimento flagelar e motilidade do espermatozoide.

2.4. QUALIDADE ESPERMÁTICA

A qualidade seminal pode ser entendida de duas formas através das características macroscópicas que estão diretamente ligadas a aspectos dos físicos, como: volume, cor, aspectos/consistência e pH, e mediante as características microscopias: motilidade, vigor, concentração e morfologia normal, que estão relacionadas com as análises mais aprofundadas da qualidade espermática, sendo estas últimas as que são normalmente utilizadas.

Assim, pode-se entender as características microscopias de motilidade e vigor como aquelas responsáveis por avaliar a movimentação dos espermatozoides, sendo que motilidade consiste em avaliar o percentual de células moveis, enquanto o vigor avaliar a cinética do movimento espermático (GARNER, 2006). Para que, as análises possam ser realizadas de maneira satisfatória nas duas características mencionadas acima é estabelecida uma diferenciação, tanto para motilidade, quanto para vigor as serão brevemente descritas abaixo.

Motilidade é compreendida em uma escala entre 0 e 100 dos espermatozoides coletados, sendo que o valor 0 refere-se aos espermatozoides imóveis e os que apresentam valores maiores que 0 são todos considerados como moveis (ZAMBELLI; CUNTO, 2006). Já o vigor determina sua escala entre 0 e 5 em suas análises, sendo que os espermatozoides com categorização igual 0 são considerados sem atividade móvel. Porém, aqueles que apresentam categorização maior que 0 são classificados com atividade móvel com seu respectivo grau de movimentação, isto é, o vigor avalia não só a movimentação como um todo, mas sim a frequência em que essas movimentações ocorrem, assim essas movimentações podem ser classificadas da

seguinte forma: movimentações leves, vibrações, movimentos para frente moderado e movimentos para frente muito ativos.

Assim, a qualidade de amostras seminal estuda e determina quais parâmetros têm percentual de espermatozoides normais e anormais em uma amostra seminal e quais os impactos relacionados a sua fertilidade (BLOOM, 1973), classificando as alterações espermáticas em defeitos maiores e defeitos menores. Os defeitos maiores estão relacionados com a infertilidade, patologias testiculares ou epididimárias, enquanto que os defeitos menores são referentes as anomalias de menor impacto na fertilidade como os erros durante a manipulação do ejaculado, ou melhor, há um limite estabelecido para o percentual de defeitos totais, os defeitos maiores não devem ultrapassar 0.10 e os defeitos menores não devem exceder 0.20 (CBRA, 2013).

Outro ponto importante sobre a qualidade seminal é sua relação com as características da integridade da membrana plasmática dos espermatozoides, a qual é fundamental para manutenção da homeostase celular, pois está diretamente ligada a capacidade de fértil dos animais (ANDRADE et. al., 2007). Caso venha ocorrer uma lesão na membrana isso significa que os espermatozoides não tem integridade suficiente (MOCE; GRAHAM, 2008).

SEÇÃO 3

PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Nesta seção, são apresentados alguns conceitos básicos sobre o processo de Descoberta de Conhecimento em Bases de Dados, bem como a tarefa de mineração de dados e suas técnicas, descrevendo os algoritmos *Apriori* e o *FP-Growth* e suas aplicações.

3.1. MINERAÇÃO DE DADOS E A DESCOBERTA DE CONHECIMENTO

A descoberta de conhecimento em bases de dados – KDD, nada mais é que um conjunto de etapas ou fases correlacionadas de maneira dependente, cujo, seu objetivo incomum é detectar padrões em bases de dados transacionais. Esse processo de detecção usa a tarefa de mineração de dados como fase principal dessa busca por padrões. Outras áreas de conhecimento como a estatística, matemática convencional e a inteligência artificial, adotam métodos diferentes para a identificação desses padrões contidos nos dados.

Segundo (ELMASRI, 2005, 625p) a descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases*), normalmente é abreviada como KDD, engloba mais que a mineração de dados. O processo é composto por seis fases: seleção de dados, limpeza, enriquecimento, transformação ou codificação, mineração de dados e construção de relatórios, e apresentação da informação descoberta.

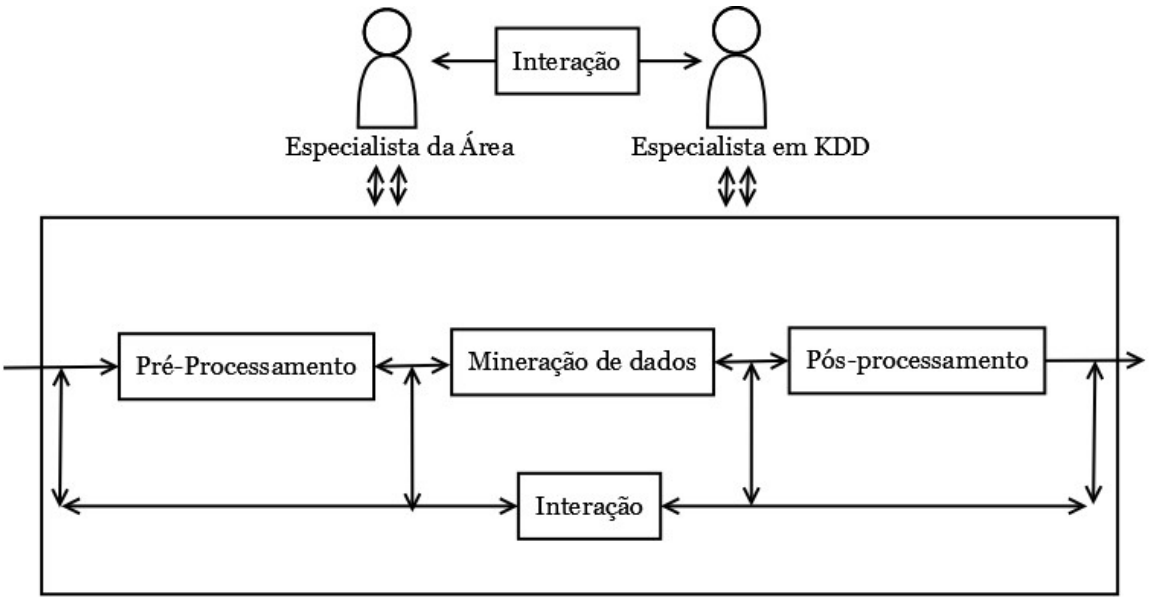
Para que a aplicação de uma tarefa de extração de dados seja bem sucedida é necessário basicamente três componentes em sua aplicação, sendo eles:

- **Conjunto de dados**: que como o próprio nome sugere, é um conjunto de dados armazenados em determinado local, podendo está organizado ou não, os quais posteriormente serão utilizados no processo de KDD.
- **Especialista da área**: o especialista do domínio da aplicação é a pessoa que tem conhecimento específico sobre o assunto que será aplicado.

- **Especialista em KDD**: é a pessoa do grupo que tem conhecimento e a experiência na implementação e aplicação de técnicas do próprio KDD.

Segundo (GOLDSCHMIDT et. al., 2015, 22p) foi proposto um ciclo de etapas operacionais do processo de descoberta de conhecimento em bases de dados conforme a Figura 2.

Figura 2. Etapas Operacionais do processo de KDD



Extraído de (GOLDSCHMIDT et al. 2015).

Dessa forma, o KDD é um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (GOLDSCHMIDT et. al., 2015, 4p). Sendo assim, a complexidade no processo de KDD consiste em perceber, observar e interpretar adequadamente cada iteração que compõe o mesmo na busca de novos conhecimentos úteis. Para que isso ocorra adequadamente temos a etapa de pré-processamento que é responsável por selecionar, limpar, enriquecer e transformar os dados em formatos compreensíveis que serão utilizados pela etapa de mineração de dados, pois é a partir desta etapa que a detecção de padrões propriamente dita acontece. Consequentemente a essa fase temos a etapa final conhecida como pós-processamento onde são feitas as análises e as interpretações dos resultados obtidos com objetivo de utilizar as informações descobertas na tomada de decisões.

3.2. PRÉ-PROCESSAMENTO DE DADOS

O pré-processamento compreende todas as funções relacionadas com a captação, organização e o tratamento de dados (GOLDSCHMIDT et al., 2015, 23p). Esta etapa tem como objetivo a preparação dos dados para a etapa de mineração de dados. Nesse sentido, compete ao pré-processamento transformar os dados em parâmetros de entrada com formato apropriado para análises subsequentes (TAN et al. 2009, 4p). Segundo o mesmo autor, a limpeza de dados está diretamente ligada a remoção de ruídos e a remoção de dados duplicados, no entanto, essa seleção de dados inconsistentes deve ser tratada com maior cuidado, pois é através dessa atividade que os resultados não terão analises equivocadas.

Devido ao pré-processamento ser responsável pela coleta, verificação de dados duplicados e/ou faltosos, dentre outras inconsistências que os dados podem apresentar essa etapa requer um maior esforço do especialista em KDD. Pois, qualquer modificação incorreta que venha a ser realizada na base de dados pode fazer como que os dados percam sua integridade assim é necessário validar qualquer modificação com os especialistas da área.

3.3. SELEÇÃO DE DADOS

É o estágio ou função considerado como redução de dados que inclui a escolha e seleção dos dados nas próprias bases de dados pelos especialistas da área e pelos especialistas em KDD. Sendo assim, os dados processados nessa etapa serão utilizados durante todo o processo de análise e de aplicação da descoberta de conhecimento. Mediante a essa escolha e seleção de dados tenta-se verificar se há algum problema na maneira em que os dados estão alocados.

De acordo com (TAN et. al., 2009, 6p), os dados necessários para uma avaliação podem não estar armazenados em um único local, ou não são de propriedade de uma organização, ao invés disso, eles estão distribuídos em bases de dados heterogêneas, isto é, para cada domínio o esforço na seleção de dados será diferente. Além disso, a tarefa de seleção de dados é crítica, porque os dados podem não estar disponíveis em um formato apropriado para serem utilizados no processo de KDD (ALMEIDA & DUMONTIER, 1996).

3.4. LIMPEZA DE DADOS

A limpeza de dados tem a função de garantir a qualidade dos dados mesmo que não haja qualquer alteração dos tipos: enriquecimento de dados, tratamento de ruídos, campos em brancos, dentre outras anomalias que podem vim a ocorrer. Uma vez que sejam mantidas as características de qualidade como: integridade, completude e veracidade, pode-se assegurar que as etapas futuras de mineração de dados não irão processar informações equivocadas e/ou irrelevantes que serão futuramente utilizadas pela etapa de pós-processamento assim evitando a produção de resultados errados.

Esta atividade abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade, integridade) dos dados representados (GOLDSCHMIDT et. al., 2015, 23p). Vale ressaltar que nenhuma base de dados possui dados 100% íntegros. Nelas podem haver inconsistências, como: erros humanos, limitação dos dispositivos de armazenamento ou das tecnologias utilizadas e a falta de sincronização, dentre outras.

Uma subatividade da limpeza de dados é a de eliminar dados estranhos ou dados em branco, pertinentes em uma base de dados, ou seja, dados que não tem nenhuma utilidade ou tem muitos valores faltosos, pois uma análise confiável pode ficar difícil ou inviável. Conforme (TAN et. al., 2009, 48p), é uma estratégia simples e eficaz, é eliminar objetos com valores ausentes, entretanto, mesmo que um campo ou um atributo da base de dados estejam com valores faltosos é preciso que o especialista em KDD tenha o cuidado de saber se aquele item não armazena nenhum tipo de informação útil. Porém, se muitos dados estiverem ausentes na base de dados essa análise pode não ser confiável ou ter mesmo inviável.

Então, independente dos dados estarem armazenados em bases de dados centralizadas ou em bases de dados distribuídas, é preciso que os dados apresentem o mínimo possível de consistência e integridade para que o problema proposto possa ser solucionado de maneira eficaz e eficiente, pois é a partir dessas métricas que os resultados obtidos terão uma qualidade aceitável e as interpretações finais serão resolutivas para o problema.

3.5. CODIFICAÇÃO OU TRANSFORMAÇÃO DE DADOS

Normalmente os algoritmos de mineração de dados requerem que os formatos naturais em que os dados se encontram sejam codificados ou transformados em formatos compreensíveis pelos algoritmos de mineração, ou melhor, essa transformação deve garantir que os dados estejam em formato adequado e/ou pronto para serem minerados. A seguir algumas definições de codificação de dados, conforme (GOLDSCHMIDT et. al., 2015):

- **Construção de atributos:** esta operação consiste em gerar novos atributos a partir de atributos existentes. Os novos atributos são determinados **atributos derivados.** Como exemplo pode-se citar a criação de atributo “idade” a partir do atributo “data_nasc” (data nascimento) e da data corrente.
- **Enriquecimento de dados:** a função de enriquecimento consiste em agregar mais informações a cada registro do conjunto de dados, para que estes forneçam mais elementos para o processo de descoberta de conhecimento.
- **Normalização de dados:** esta operação consiste na escala dos valores de cada atributo de forma que estes sejam mapeados para valores restritos a pequenos intervalos, tais como de -1 a 1 ou de 0 a 1. Tal ajuste faz-se necessário para evitar que alguns atributos, por apresentarem uma escala maior que outros, influenciam de forma tendenciosa em determinados métodos de mineração de dados.

3.6. MINERAÇÃO DE DADOS

A mineração de dados é uma parte integral da descoberta de conhecimento em bancos de dados (*KDD - Knowledge Discovery in Databases*), na qual é o processo geral de conversão de dados brutos em informações úteis. (TAN et. al., 2009, 4p). Sendo assim, a mineração de dados é considerada como a etapa de maior relevância no processo de descoberta de conhecimento, pois ela é responsável pelo processamento dos dados selecionados e tratados em atividades anteriores. Com isso a mineração de dados pode ser entendida como uma atividade multidisciplinar, porque sua aplicação vária de acordo com a área de atuação.

Com uma perspectiva voltada a área de estatística temos a definição de (HAND et. al., 2001), na qual trata a mineração de dados como aquela que é responsável por analisar grandes conjuntos de dados, a fim de encontrar relacionamentos inesperados, assim ela resume os dados de uma forma que eles sejam tanto úteis, quanto compreensíveis aos donos da base de dados.

Outra definição dada por (FAYYAD et. al., 1996) é que, Mineração de Dados é um passo no processo de descoberta de conhecimento, que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados.

Mediante ao que foi exposto acima, a mineração de dados pode ser um processo semi-automatizado ou automatizado que tem o propósito de identificar padrões contidos em bases de dados independentemente da sua origem, mesmo que esses dados possam estar armazenados em áreas geográficas distintas. Com isso, a mineração de dados preocupa-se com aquisição de conhecimento útil que possa ser extraído dessas fontes de dados, a qual tenta encontrar informações relevantes que permanecem ignoradas.

No entanto, nem todas as tarefas de descoberta de conhecimento são interpretadas como mineração de dados. Por exemplo: a procura de registros individuais usando um sistema gerenciador de banco de dados ou pesquisar por páginas webs através de sites de busca, fazendo uso da internet. Essas atividades são consideradas como recuperação de dados.

Na fase da aplicação de mineração de dados, faz-se necessário identificar quais tarefas e técnicas farão parte desse processo para que a atividade de descoberta de conhecimento ocorra de maneira eficiente, ou seja, as tarefas e as técnicas devem ser definidas previamente levando em consideração as características descritas pelos especialistas área a serem resolvidas. A seguir serão apresentadas as tarefas mais comuns da etapa de mineração de dados:

- Associação: as regras de associação representam padrões onde a ocorrência de eventos em um conjunto de dados é alta, isto é, as regras de associação têm a finalidade de encontrar itens que estão frequentemente relacionados em transações dos dados armazenados indicando que sempre que há ocorrência de um também terá ocorrência de outro na maioria das vezes. Por exemplo,

considere que um cliente A compra um item X da prateleira e o mesmo também costuma levar um item Y, fazendo relação a estrutura (SE ENTÃO), se X for comprado então Y também será comprado.

- Classificação: a classificação é a tarefa mais conhecida e importante do processo do KDD, assim a classificação pode ser definida como um processo supervisionado que é um aprendizado indutivo, ou seja, um aprendizado capaz de relacionar eventuais atributos existentes entre uma base de dados. Para que isso aconteça os atributos desta tarefa são divididos em dois grupos, sendo que um têm características comuns, ou melhor, contém atributos previsores ou de previsão. Já os outros contém apenas um único atributo que é denominado de atributo alvo ou atributo da classe o qual pode classificar e analisar os dados pertencem a uma classe ou não.
- *Clustering* ou (agrupamento): consiste em separar os registros de uma base de dados em subconjuntos ou grupos (*cluster*), de maneira que os atributos dos grupos estejam relacionados às características comuns entre todos que os diferencie de outros clusters. Para (ELMASRI, 2005, 637p) “o objetivo do agrupamento é colocar os registros em grupos, de tal forma que os registros de um grupo sejam similares aos demais do mesmo grupo e diferentes daqueles dos demais grupos”.
- Regressão: a tarefa de regressão é constituída a partir de uma modelagem preditiva onde sua variável alvo é compreendida como continua, ou seja, para que um problema X qualquer possa ser resolvido é necessário usar outros indicadores que se relacionam a este problema para que ele seja solucionado. Por exemplo: a previsão de vendas de uma empresa é baseada na quantidade de publicidade que a mesma produz.

3.7. PÓS-PROCESSAMENTO E INTERPRETAÇÃO DE DADOS

Esta etapa do processo de descoberta de conhecimento do KDD envolve as técnicas de análise, avaliação e de visualização dos resultados obtidos através da etapa de mineração de dados. É, a partir dessa etapa de descoberta de conhecimento que os especialistas em KDD juntamente com os especialistas da área avaliam e analisam os resultados encontrados para comprovar ou não paradigmas referentes

aos problemas, essa comprovação é feita através de critérios, como: grau de confiança da regra e suporte.

Tendo em vista os resultados apresentados e obtidos pela etapa de mineração de dados várias dessas informações são expressas de maneira complexa dificultando o seu entendimento. Então, é necessário utilizar técnicas e/ou métodos de transformação de dados para facilitar essas análises, pois só assim pode haver uma melhor compreensão do que está exposto nos resultados com um grau de confiabilidade aceitável, pois assim o usuário final poderá fazer uso das informações oferecidas pela etapa de mineração de dados.

3.8. REGRAS DE ASSOCIAÇÃO

A tarefas de associação tem como objetivo buscar correlação entre atributos em uma base de dados, em outras palavras busca encontrar itens que impliquem na presença de outros na mesma transação. Por exemplo: são comumente utilizadas para a identificação de padrões em transações de compras conjuntas de supermercados. Pois, as regras de associação permitem que os especialistas em KDD tenham uma visão mais clara dos itens que se apresentam com maior frequência em bases de dados discriminando novos padrões. Portanto, as regras de associação são as mais adequadas para o problema abordado neste trabalho pois elas buscam identificar padrões: de aspectos físicos e de qualidade seminal através da relação entre os dados coletados com o período reprodutivo dos macacos-de-cheiro que vivem em cativeiro.

A implicação dessa tarefa consiste em determinar que um dado item X e um dado item Y, tem relação implícita ou explicita em um conjunto de dados onde a ocorrência X determina que haja a ocorrência de Y expressos pela estrutura lógica (SE X ENTÃO Y). Sendo assim, na maioria das vezes que X for verdadeiro Y também será verdadeiro sendo X o antecessor e Y sucessor da expressão.

Segundo (TAN et. al., 2009, 392p) uma regra de associação é uma expressão de implicação no formato $X \rightarrow Y$, onde X e Y são conjuntos disjuntos de itens, i.e., $X \cap Y = \emptyset$. Independentemente de expressarem associação simples ou ordenadas e/ou coordenada, são comumente representadas por meio das afirmações do tipo SE

ENTÃO, sendo também interpretadas como implicações do antecedente da regra (ou premissa) e consequente de regra (ou conclusão) (Silva et. al., 2016, 392p).

Uma regra de associação pode ser expressa da seguinte maneira referindo-se aos parâmetros de aspecto físicos animais estudados neste trabalho em uma transação de dados. Por exemplo: A regra (1) indica que o atributo peso aumentado (X = {peso}) pode estabelecer que os macacos-de-cheiro estão no período reprodutivo "*fatted*" (Y = {"*fatted*"}). Segundo a regra (2), a relação entre as dobras cutâneas do braço e tórax aumentadas (X = {braço, tórax}) podem induzir que os animais estão no período de reprodução (Y= {"*fatted*"}).

As regras de associação só podem ser válidas se o número de vezes da ocorrência de (X → Y) forem maiores e/ou iguais que as métricas de suporte mínimo (s_{min}) e do grau confiança mínimo (c_{min}) previamente estabelecidas, por exemplo: suporte da regra deve ≥ s_{min} e o grau de confiança da regra deve ser ≥ c_{min}. O suporte estabelece que as regras de associação devem determinar a frequência na qual a regra será aplicada, ou seja, o suporte identifica quais associações apresentam-se em quantidade expressiva em relação a outras, assim o suporte é definido como *itemset* X U Y divididos pelo número de transações como mostra a Equação (a). Por outro lado, o grau de confiança delimita e contabiliza a quantidade de vezes que o atribuo sucessor ocorre em uma transação que tenha sempre o mesmo antecessor conforme Equação (b).

$$\text{Suporte, s } (X \rightarrow Y) = \frac{\sigma(X \cup Y)}{n} \qquad \text{(a)}$$

$$\text{Confiança, c } (X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \qquad \text{(b)}$$

Os dados dos parâmetros de aspectos físicos (braço, peso, tórax e volume testicular) aumentados, podem ser representados em um formato binário na Tabela 1, onde cada linha corresponde a uma transação e cada coluna refere-se a um item. Um item pode ser tratado como uma variável binaria, para os valores presentes na transação usa-se o número um e para os não presentes utiliza-se zero.

Tabela 1. Uma representação binária dos dados de aspectos físicos aumentados

TID	Braço	Peso	Tórax	Vol. Test	Período
1	1	1	0	0	1
2	1	1	1	0	1
3	0	0	1	1	0
4	1	1	0	1	1
5	0	0	1	0	1

Fonte: Arquivo pessoal, (2017)

As aplicações prática das equações citadas anteriormente serão apresentadas a seguir: “Um macaco que esteja com as medidas do Braço e Peso aumentadas Tabela 1 {Braço, Peso} indicam que os macacos estão no período reprodutivo {”*Fattef*”}, a resolução da regra da descoberta expressa-se da seguinte forma ({Braço, Peso} → {”*Fatted*”}). Sendo que o contado de suporte para {Braço, Peso, ”*Fatted*”} que é 3 e o número total de transação é igual a 5 assim o suporte da regra é 3/5 = 0,6. O grau de confiança utiliza um contador de suporte dos itens de {Braço, Peso e ”*Fatted*”} que é dividido pela quantidade de ocorrências de {Braço, Peso}, onde a quantidade de ocorrências {Braço, Peso e ”*Fatted*”} é 3 e as ocorrências de {Braço, Peso} é 3 resultando no seguinte grau de confiança da regra: 3/3 = 1.

Vale ressaltar que as métricas de suporte e do grau de confiança são de extrema importância na descoberta de padrões, sendo que cada uma trata e extraí diferentes informações das transações de banco de dados. O suporte em sua aplicação incorreta pode ocasionar regras com baixo índice de relevância a partir de uma perspectiva de negócio, ou melhor, regras com baixo índice de suporte podem promover determinados itens que não sejam realmente os de maior relevância ou problema em questão quando adquiridos em conjunto. Por outro lado, o grau de confiança tem o papel de medir a confiabilidade da inferência realizada em uma regra, ou seja, uma determina regra X → Y só é considerada válida, quanto maior for seu grau de confiança, logo a probabilidade de Y está presente em uma transação que possua o X como antecessor será maior.

Então, as regras de associação utilizam algumas estratégias de tratamento dos dados de uma transação uma dessas estratégias seria decompor o problema em duas

subtarefas tais como: geração de conjuntos de itens frequentes e a geração de regras. Geração de conjuntos de itens frequentes tem o objetivo de encontrar todos os conjuntos de itens que satisfaçam o limite do suporte, já a geração de regras tem a finalidade de extrair todas as regras com alto grau confiança dos conjuntos de itens frequentes (TAN et. al., 2009, 394p).

3.8.1. ALGORITMO APRIORI

O *Apriori* é um dos algoritmos clássicos usados em mineração de dados para resolução de problemas que utilizam tarefas de associação na descoberta de conhecimento em bases de dados, sua aplicação baseia-se no suporte para controlar de forma sistemática o crescimento exponencial dos conjuntos de itens candidatos (TAN et. al., 2009, 399p). O algoritmo *Apriori* é composto por duas etapas a primeira etapa é responsável por identificar a quantidade de repetição que cada item possui (ou *itemset)* em uma base de dados transacional. Em seguida, controla os itens com o suporte maior que o mínimo estabelecido pelo usuário sendo que eles são selecionados e combinados para compor *2-itemset*s, *3-itemset*s, 4-*itemsets* e assim por diante até que não haja mais a produção de *itemsets* frequentes (Silva et. al., 2016, 208p).

Assim, o *Apriori* aplica um contador de iteração para cada processo *k,* estabelece um conjunto de transações T_{ID} *e* um suporte mínimo S_{min} para cada *itemsets, e*sses valores servem como entradas no algoritmo *Apriori* até que não haja mais a produção de *itemsets* frequentes no processo de descoberta de conhecimento. Assim, para cada iteração é criado e calculado o seu S_{min} para todo os *itemset* presentes em uma $T_{ID,}$ sendo essa a primeira atividade da execução do algoritmo em seguida os itens são inseridos em uma lista (L_K). Porém, apenas *os itemsets* maiores que S_{min} serão inseridos nessa *Lk* e este processo ocorrerá até que nenhum *itemset* seja mais encontrado conforme o Algoritmo 1 (Silva et. al., 2016, 208p).

Algoritmo 1: Primeira fase do algoritmo *Apriori*: identificação dos itemsets frequentes

Parâmetros de entrada

- T_{ID}: um conjunto de transações;
- S_{min}: suporte mínimo;
- ***K***: contado inicializado em 1;

Parâmetros de saída

- ***L***: lista de *itemsets* frequentes;

Passo 1: calcular o suporte de cada 1 –*itemset* de T_{ID};

Passo 2: selecione os 1 –*itemsets* com suporte maior ou igual a s_{min} *e os insira* em L_K;

Passo 3: **enquanto** L_k <> 0 faça

Passo 3.1: combine os *itemset* gerando o conjunto de (k + 1) –*itemsets* candidatos;

Passo 3.2: calcule o suporte de cada novo *itemset*;

Passo 3.3: k++;

Passo 3.4: selecione os *itemsets* com suporte maior ou igual a s_{min} *e os insira* em L_K;

Passo 3.5: ***L*** U L_K;

Fonte: Extraído de (Silva, 2016).

Vale ressaltar que durante a aplicação do algoritmo *Apriori* ocorre a poda de alguns *itemsets* que nada mais é que a remoção dos *itemsets* com baixo índice de incidência em um processo transacional, onde o seu número *k* é menor frente aos de outros candidatos (ou *itemsets*). Segundo (TAN et. al., 2016, 402p), o método de força bruta analisa cada conjunto de *k* de itens como um potencial candidato e depois aplica o passo de poda dos candidatos para remover qualquer candidato desnecessário.

Para realizar a poda através da métrica do grau de confiança utiliza-se o teorema ilustrado pelo mesmo autor citado anteriormente, o qual determina que se uma regra X → Y-X não satisfaça o limite de confiança da regra, então, qualquer regra X’ → Y’-X’, onde X’ é um subconjunto de X também não deverá satisfazer o limite do grau de confiança da regra. No entanto, mesmo as regras que tenham o grau de

confiança dentro de intervalo estipulado podem sofrer poda desde que as mesmas possuam apenas 1 –*itemset* em todo processo k.

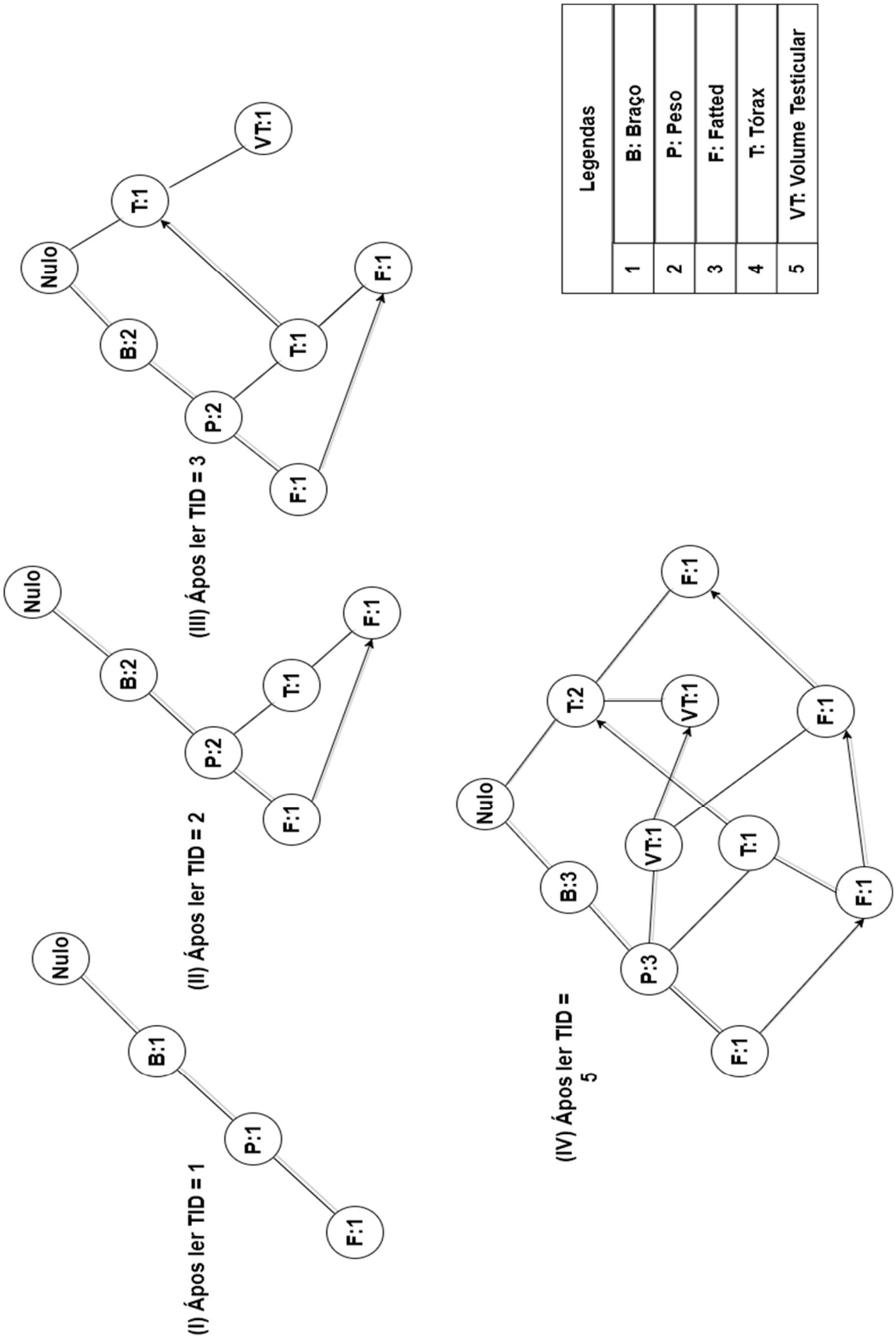
3.8.2. ALGORITMO *FP-GROWTH*

As tarefas de associação tradicionais abordam e assumem técnicas próximas as do algoritmo *Apriori*, o qual baseia-se na procura de itens frequentes em um comprimento (*k* + 1) de uma transação de dados, caso existam itens frequentes essa procura por *itemsets* em um comprimento *k* torna-se custosa e demanda bastante tempo, pois a quantidade de repetições realizadas em uma busca por itens candidatos tem um alto custo em termos computacionais. Para (Han et. al., 2004) é dispendioso percorrer repetidas vezes uma base de dados para verificar e testar todos os conjuntos de itens candidatos e seus padrões correspondentes.

Então, alguns métodos alternativos têm sido desenvolvidos para suprir e melhorar tais eventualidades com o objetivo de melhorar o processo de busca de padrões através das regras de associações, assim outros algoritmos foram desenvolvidos. Por exemplo: *DHP (Direct Hasbing and pruning), Partition, Eclat,* dentre outros, sendo todos eles inspirados no *Apriori.* Porém, houve um algoritmo alternativo que organiza sua base de dados em estruturas hierárquicas em forma de árvores diferindo expressivamente da abordagem do algoritmo *Apriori* que gera e testa *itemsets,* esse algoritmo é chamado *FP-Growth – (Frequente-Partten Growth)*. O *FP-Growth* é um algoritmo que não concorda com o paradigma gerar e testar do *Apriori*, ao invés disso, ele codifica o conjunto de dados usando uma estrutura de dados compacta chamada árvore- FP (TAN, et. al., 2009, 433p).

Uma árvore-FP é uma representação hierárquica de um conjunto de dados de uma transação assim mapeando cada uma dessas transações em caminhos de uma árvore-FP, ou seja, esses caminhos criados fazem o cruzamento entre os itens e esses cruzamentos são os *itemsets* frequentes das transações. Essa forma de organizar as transações em árvores faz com que o uso de memória principal seja reduzido dando agilidade e eficiência ao processo de descoberta de padrões.

Figura 3. Construção de uma árvore-FP



Fonte: Arquivo pessoal (2017)

A Figura 3 mostra o conjunto de dados de dados da Tabela 1 que contém 5 transações e 5 itens. Após a leitura das três primeiras transações é criada uma árvore–FP corresponde para cada uma das transações lidas, assim cada nodo (ou ramo) criado na árvore-FP contém todos itens mapeados e organizados para um determinado caminho até que não haja mais itens a serem inseridos e árvore-FP seja completada. Vale ressaltar, que antes de tudo a árvore-FP apresenta apenas único nó nulo que é o seu nó raiz.

A primeira transação é composta pelos itens {B, P, F} criando um ramo que inicia do nó raiz e vai até o último item da transação lida, essa leitura dos itens cria caminhos da seguinte forma nulo → B→ P→ F tendo sempre um contador para cada item igual a 1. A Segunda transação é constituída pelos itens {B, P, T, F} e também é iniciada no nó raiz sua construção é expressa a seguir nulo → B → P → T → F. Porém, essa transação possui 2 nós incomum com a primeira transação esse ponto em incomum (ou sobreposto) faz com que o contador frequência do item passe de 1 para 2 indicando que naquele ponto há a incidência de itens frequentes (ou *itemsets*) e também apresenta 1 nó adjunto {F} com a primeira transação. A leitura da terceira transação {T, VT} forma um novo ramo que também é iniciada no nó raiz e expressa da seguinte maneira nulo → T → VT interligando todos itens da transação com o contador de frequência igual a 1. No entanto, assim como a segunda transação a terceira apresenta um nó adjunto {T} com a segunda transação. Após todo esse processo temos a última transação com todas as transações e seus respectivos itens e caminhos mapeados em forma de árvore.

Vale ressaltar, que algoritmo *FP-Growth* estrutura suas transações em forma de árvores e a leitura dessa árvore criada pelo *FP-Growth* se dá de maneira *bottom-up* de baixo para cima para encontrar seus os *itemsets*, todos os caminhos gerados adotam a estratégia dividir-para-conquista, onde o problema é fracionado em subproblemas para melhorar o desempenho de busca. Dessa forma, o algoritmo *FP-Growth* preocupa-se apenas em fazer a leitura dos itens que apresentam um alto grau de incidência.

3.9. TRABALHO CORRELATO

Há vários trabalhos correlatos referentes ao processo de descoberta de conhecimento em bases de dados que utilizam das mais diversas metodologias que englobam o processo de KDD. No entanto, para aplicação de Medicina Veterinária voltada especificamente à análise de parâmetros físicos e morfológicos apenas um trabalho foi encontrado, o qual é brevemente descrito a seguir.

A “Descoberta de Padrões no Sêmen de Primatas Não-Humanos através da Mineração de dados”, realizado pelas pesquisadoras (Aline et. al, 2013) junto ao BIOMEDAM – (Laboratório de Biologia e Medicina de Animais Silvestres da Amazônia) que teve o objetivo de utilizar a Mineração de dados para detectar padrões nas características testiculares de primatas não-humanos, correlacionado com a qualidade do sêmen. Para essa pesquisa foram utilizadas as técnicas de associação e classificação de dados para identificação de padrões e posteriormente foi feita uma comparação dos resultados obtidos pelas técnicas. Após esse processo de testes com os dados foram obtidos os seguintes resultados: circunferências testiculares: reduzidas e normais podem produzir sêmen em quantidades normais e de boa qualidade; produção de espermatozoides com melhor não está relacionada a quantidade de sêmen.

Assim, o trabalho mencionado neste tópico buscou identificar padrões correlacionados as características seminais e testiculares dos primatas não-humanos por meio da mineração de dados. No entanto, A pesquisa deste trabalho buscou identificar padrões referentes ao ciclo reprodutivo dos *Saimiris Collinsi* que vivem em cativeiro, bem como correlacionar o ciclo reprodutivo com a qualidade seminal até então não identificadas essa detecção ocorreu através das regras de associação esses analises tornaram-se mais rápidas e menos complexas que as técnicas não-automatizadas.

SEÇÃO 4

ESTUDO DE CASO

Nessa seção, são apresentadas as fases de preparação dos dados para aplicação das regras de associação em um problema real. Além de descrever cada padrão encontrado através testes realizados nos parâmetros físicos e morfológicos dos macacos-de-cheiro que vivem em cativeiro.

4.1. ABORDAGEM DO PROBLEMA E OS ATRIBUTOS UTILIZADOS

Os dados selecionados para aplicação desta pesquisa são oriundos de uma única base de dados, disponibilizada pelos pesquisadores da Universidade Federal do Pará – campus Castanhal, do curso de Medicina Veterinária, os quais distinguem suas coletas em dois momentos: o primeiro na coleta dos dados relacionados aos parâmetros de aspectos físicos e o segundo momento da coleta está relacionado aos parâmetros de qualidade seminal.

A implementação do processo de KDD para a descoberta de padrões nos dados dos macacos da espécie *Saimiri Collinsi* (macacos-de-cheiro) que vivem em cativeiro, foram realizados sob um conjunto de 560 registros de aspectos físicos e um conjunto de 1260 registros de qualidade seminal. Esses dados são organizados da seguinte maneira: 4 atributos de aspecto físico (dobras cutâneas do braço e tórax, peso e o volume testicular) e 9 atributos de qualidade seminal (concentração, grau de coagulação, motilidade, morfologia normal, vigor, volume, defeitos maiores e defeitos menores). Todos os parâmetros citados acima possuem 140 registros cada normalizados e reajustados.

Consequentemente, todos os dados coletados percorreram por todas as etapas que compõem o pré-processamento do KDD, para que as análises futuras tenham coerência e clareza em seus resultados. Após a etapa do pré-processamento originou–se 2 novos conjuntos de dados para o treinamento dos algoritmos. O conjunto 1 refere-se aos parâmetros de aspectos físicos sendo composto 4 atributos respectivamente iguais aos atributos originais, cada atributo contém 1.500 novas

instâncias, totalizando (4 x 1.500) = 6.000 novas instâncias balanceadas e ajustadas de acordo com os dados coletados. E, o mesmo processo de criação e balanceamento dos dados foi realizado no conjunto 2 de qualidade seminal sendo criado 9 atributos também iguais aos originais, onde cada novo atributo contém 1.500 novas instâncias balanceados e ajustados, totalizando (9 x 1.500) = 13.500 novas instâncias ao todo.

Devido à quantidade de dados disponíveis ser reduzida para o treinamento dos algoritmos, foi necessário criar dados sintéticos através da densidade de probabilidade de uma distribuição normal ou *Gauss* conforme a Equação (c), sendo que essa distribuição normal consiste em utilizar as métricas da média aritmética e o desvio padrão do conjunto de dados para gerar os dados sintéticos. Logo após, a esse processo de definição foram adotadas as nomenclaturas de dados reais (DR) para os registros coletados e de dados sintéticos (DS) para os dados gerados.

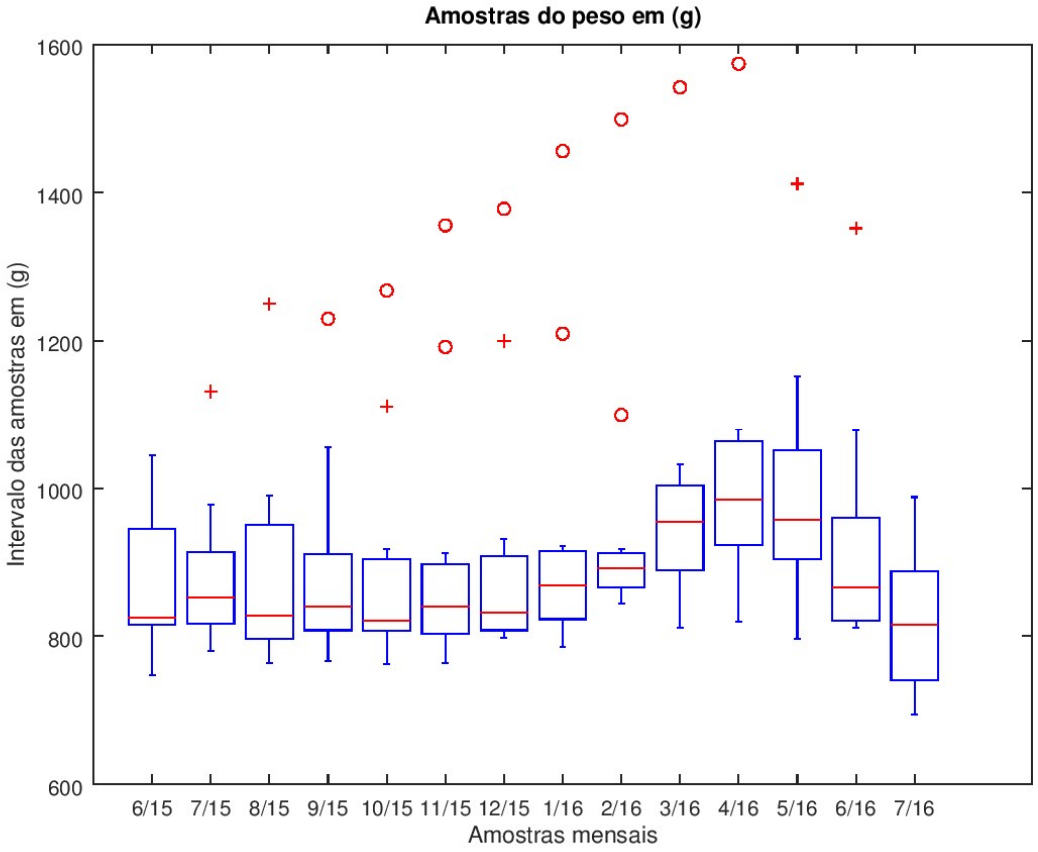
$$f(x) = \frac{1}{\sqrt{2}}\, e - \frac{1}{2}\left(\frac{x-\mu}{\quad}\right)^2 \quad \text{(c)}$$

A distribuição normal é uma das mais utilizadas em estatística em razão dela possibilitar a detecção de elementos dispersos em um conjunto de dados gerado, ou seja, dados que fogem dos limites pré-determinados pelas métricas da média aritmética e do desvio padrão por esse motivo os dados dispersos são facilmente encontrados.

4.1.1. DESCRIÇÃO DOS DADOS REAIS

Os dados usados nesse trabalho são organizados e armazenados em arquivos *XLSX* e os registros contidos nesse documento são fruto de uma pesquisa teve início em 06/2015 e foi até 07/2016, totalizando 14 meses de duração. Algumas características foram observadas no decorrer dessa pesquisa como: comportamento dos macacos, alimentação e o estresse, que os animais sofriam em cativeiro. Desse modo, as coletas eram realizadas mensalmente sob um conjunto de 10 animais na tentativa de identificar qual seria o melhor momento para colher o sêmen dos macacos. A seguir a Figura 6 mostra os dados coletados durante os 14 meses referentes ao peso dos macacos-de-cheiro.

Figura 4. Coleta dos dados do peso por meses



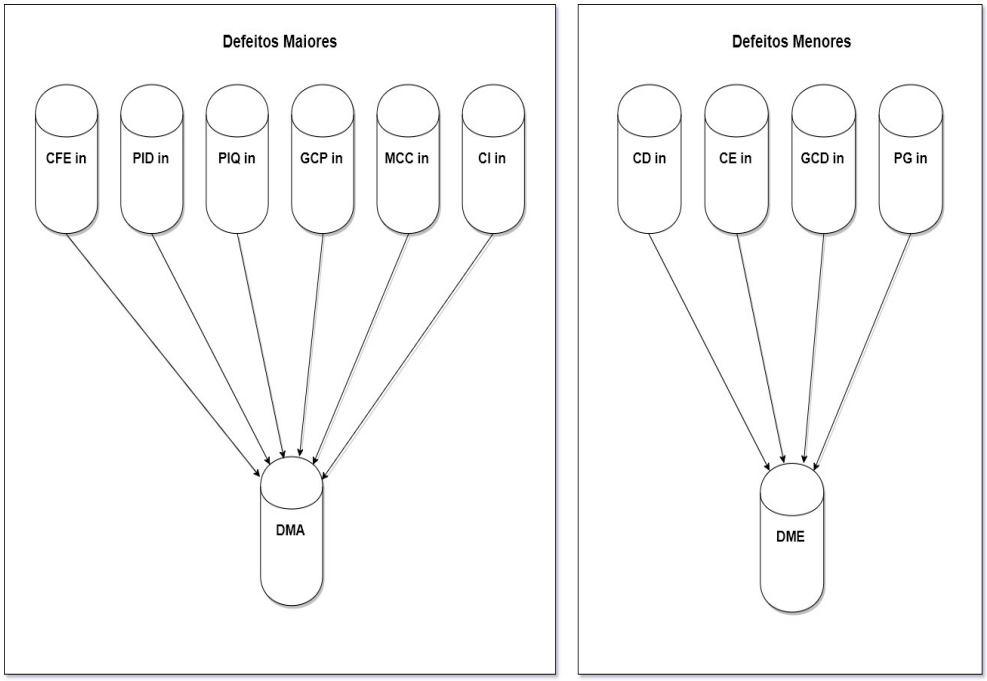
Fonte: Arquivo pessoal, 2017

A Figura 4 acima mostra o comportamento e a variação das amostras dos pesos dos macacos-de-cheiro que vivem em cativeiro, no eixo de X são apresentadas as amostras coletadas em cada mês. No eixo de Y são apresentados os limites superiores e inferiores (linha após o extremo da caixa), desvios padrões (extremos das caixas), média aritmética (linha vermelha) e os pontos discrepantes ‘o’ 3x > desvio padrão e o ‘+’ 1.5x a 3x desvio padrão (ou *outliers*) que foram detectados das amostras dos pesos, esses pontos são ditos como normais pelas análises realizadas pelos pesquisadores da área, visto que, cada animal possui uma estrutura física diferente.

Em vista disso, a base de dados trabalhada contém 29 atributos ao todo e é especificada por cada mês do ano em que a coleta foi realizada, obedecendo um período continuo. Dos atributos contidos nessa base 4 eram atributos de aspectos físicos: braço, peso, tórax e volume testicular e 25 eram de aspectos morfológicos: volume, pH, integridade da membrana plasmática, dentre outros. É importante frisar

que 2 dos atributos eram compostos a partir da soma da porcentagem de outros atributos, sendo eles: defeitos maiores e defeitos menores conforme a Figura 7.

Figura 5. Atribuição dos defeitos morfológicos



Fonte: Arquivo pessoal, 2017

A figura 7 acima mostra como os defeitos maiores e menores morfológicos são constituídos. Os defeitos maiores levam em consideração os valores de outros defeitos para sua formação, que são: Cauda fortemente dobrada – CFE, Peça intermediária dobrada – PID, Peça intermediária quebrada – PIQ, Gota citoplasmática proximal – GCP, Microcefalia – MCC, Cabeça isolada – CI. Já os defeitos menores são formados por elas seguintes métricas: Cauda dobrada – CD, Cauda enrolada – CE, Gota citoplasmática distal – GCD e Pseudogota – PG. Todos os defeitos apresentados neste trabalho são de sêmens naturais: *in nature.*

4.2. REALIZAÇÃO DO PROCESSO DE CRIAÇÃO DOS DADOS SINTÉTICOS

Os dados sintéticos foram gerados através da distribuição normal ou de Gauss como foi descrito anteriormente, para essa geração de dados foi utilizada a função normrnd da ferramenta *Octave* (Galo e Camargo, 2016) conforme a Equação (d), a qual é responsável pela geração dos dados sintéticos essa função estrutura-se da

seguinte forma: quantidades de números aleatórios a serem gerados em forma de matriz [sz], na média dos dados (μ) e no desvio padrão (σ).

$$DS = normrnd(\mu, \sigma, [sz]) \text{ (d)}$$

Consequentemente, foram utilizadas funções: *mean* e *std* da referida ferramenta. A função *mean* consiste em calcular a média aritmética dos atributos, enquanto a função *std* é responsável por calcular o valor do desvio padrão. Portanto, os valores contidos nas tabelas foram utilizados como parâmetros de criação dos dados sintéticos (DS) de treino para os parâmetros de aspectos físicos na Tabela 1 o mesmo processo foi feito para os atributos de qualidade seminal.

Tabela 2. Média e Desvio Padrão dos dados de aspectos físicos

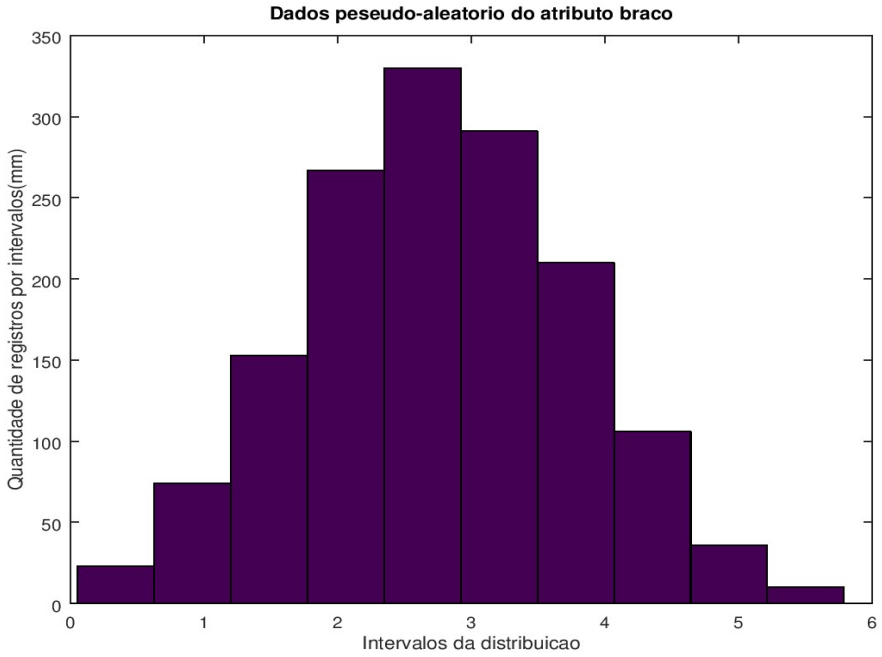
Atributos	Médias (μ)	Desvio padrão (σ)
Braço (*mm*)	2.6128	0.88536
Peso (*g*)	882.71	103.08
Tórax (*mm*)	2.2733	0.83009
Volume testicular (*cm*)	31.877	15.220

Fonte: Arquivo pessoal, 2017

Dessa forma, todo o processo de definição das médias aritméticas e dos desvios padrões dos dados é de suma importância, pois é a partir dessas métricas que os dados sintéticos são gerados. Nesse sentido, os dados sintéticos têm dependência exclusiva dessas métricas diminuído as chances de ter dados discrepantes gerados. Segundo (TAN et. al., 2009) a chance de haver um elemento estranho gerado em uma distribuição normal é reduzida, no entanto, caso haja, é minimamente discrepante ao limite determinado.

Essa detecção de elementos estranhos é facilmente perceptível, pois está diretamente relacionada aos dados sintéticos gerados conforme a Figura 4 sendo que esses dados levam em consideração as referidas métricas em sua instanciação. Pois são responsáveis por estabelecer quais limites laterais uma distribuição normal pode ter, isto é, a geração dos dados sintéticos em tese não deve ultrapassar o valor estabelecido pelo desvio padrão tendo como referencial sua respectiva média.

Figura 6. Distribuição dos dados gerados referentes medidas do braço em (mm)



Fonte: Arquivo pessoal, 2017

A Figura 4 mostra o comportamento dos dados gerados a partir de uma distribuição normal referente as métricas do peso usadas na (Tabela 1), assim é possível notar que os dados gerados se concentram em sua maioria próximos a métrica da média dos atributos. Nesse sentido, os dados de treinos obedecem aos limites determinados para que os testes com os dados reais fossem feitos de maneira eficiente. Consequentemente todos dados gerados foram normalizados e ajustados conforme os parâmetros previamente determinados no início do processo de descoberta de conhecimento.

4.3. DESCRIÇÃO DE COMO FOI REALIZADO O TRATAMENTO DOS DADOS SINTÉTICOS E REAIS

Para utilização dos dados reais e sintéticos pelos algoritmos *Apriori* e *FP-Growth*, faz-se necessária o processo de discretização dos dados, ou seja, estes foram transformados em intervalos facilmente compreensíveis para um melhor desempenho das regras de associação. Por exemplo: o atributo nominal período reprodutivo {“*no-fatted”*, ”*fatted*”} pode ser substituído por um par de itens binários, onde o período “*fatted*” passa a ser {1} e o período “*no-fatted*” igual a {0}. Como isso, o processo de transformação também foi utilizado nos conjuntos de dados deste

trabalho que serão apresentadas na Tabela 2 a qual mostra a organização dessa discretização (ou transformação) dos dados.

Tabela 3. Discretização dos dados voltada para os aspectos físicos

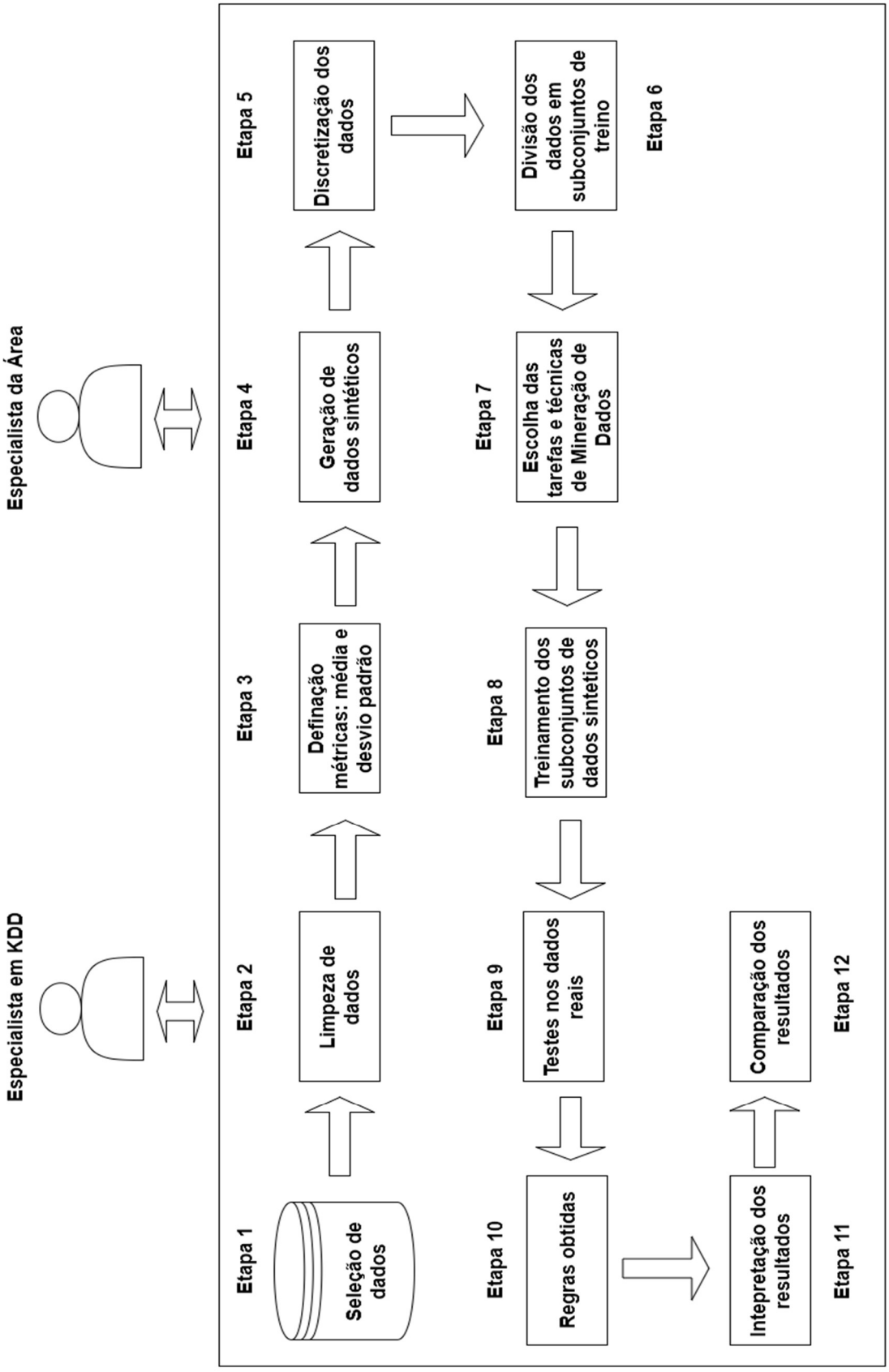
Atributos/Discretização	Baixo	Médio	Alto	Muito Alto
Braço	[-inf – 1,9]	[1,9 – 2,5]	[2,5 – 3,5]	[3,5 - inf+]
Peso	[-inf - 812]	[812 - 860]	[860 - 990]	[990 - inf+]
Tórax	[-inf – 1,5]	[1,5 – 2,0]	[2,0 – 2,9]	[2,9 - inf+]
Volume Testicular	[-inf – 1,5]	[1,5 – 2,1]	[2,1 – 3,0]	[3,0 - inf+]
Período	"*no-fatted*""		"*fatted*"	

Fonte: Arquivo pessoal, 2017

Vale ressaltar que essa discretização foi determinada mediante os períodos pré-estabelecidos pelos pesquisadores da área, os quais foram os responsáveis pela coleta dos dados num período de 14 meses corridos entre o mês 6 de 2015 ao mês 7 de 2016. Após a etapa de coleta de dados os pesquisadores definiram que os primeiros 7 meses são classificados como *'' no-fatted''* e os 7 últimos meses como *'' fatted''*, sendo esses uns dos paradigmas a serem confirmados neste trabalho.

No que diz respeito a discretização de dados é de suma importância mencionar a variável MACP – (média aritmética de cada período), a qual é responsável por calcular as médias aritméticas dos períodos "no-*fatted*" e *fatted* e armazená-las, isto é, MACP calcula as médias aritméticas de cada período de maneira isolada para que posteriormente sejam contabilizadas as quantidades de dados maiores que a MACP. Por exemplo: conjunto de dados com 40 instâncias deve conter no mínimo 28 instâncias maiores que o seu MACP correspondente isso equivale a 70% das instâncias analisadas. Logo após ao desenvolvimento dessa métrica MACP foi necessário avaliá-la juntamente com os especialistas da área, para que os resultados dessa pesquisa não se apresentassem de maneira equivocada. A seguir a Figura 7 apresenta todo o processo de manuseio dos dados deste trabalho para a descoberta de padrões.

Figura 7. Etapas do processamento de dados



Fonte: Arquivo pessoal, 2017

A seguir serão brevemente descritas as etapas realizadas neste trabalho (Figura 7): a etapa 1 é responsável por selecionar os atributos necessários; a etapa 2 consiste em garantir integridade dos dados; a etapa 3 calcula as médias e os desvios padrões dos dados; a etapa 4 consiste em gerar novas instâncias; a etapa 5 é responsável por transformar os dados em formatos compreensíveis pelos algoritmos; a etapa 6 consiste em agrupar os dados em subconjuntos; na etapa 7 é definida qual tarefa será utilizada, bem como os algoritmos; a etapa 8 faz o treinamento dos algoritmos; a etapa 9 realiza os testes dos dados reais; a etapa 10 são as regras obtidas após os testes; na etapa 11 é feita a intepretação das regras e por fim na etapa 12 é feita as análises comparativas entres os resultados de veterinária com os deste trabalho.

4.4. DESCRIÇÃO E CONVERSÃO DOS CONJUNTOS DE TREINO

Após, todo o processo de construção dos conjuntos de dados de treino ou subconjunto de dados foi necessário realizar a conversão dos dados em arquivos TXT para arquivos como a extensão ARFF sendo esse o formato padrão de leitura da ferramenta WEKA (Eibe, 2016). Assim, todos os arquivos utilizados nessa pesquisa foram transformados e convertidos para esse formato conforme a Figura 5. A referida figura apresenta uma estrutura que é composta por um cabeçalho (*@relation),* seguida pela declaração dos atributos representado notação (*@atrtibute)* com seu respectivo tipo e pôr fim a notação (*@data)* que é responsável por organizar os valores correspondentes aos atributos declarados. Todos os arquivos de leitura do WEKA seguem essa sequência de declaração para funcionar corretamente.

Figura 8. Estrutura padrão de leitura do WEKA

```
@relation reproducao

@attribute braco {bbaixo,bmedio,balto,bmalto}
@attribute peso {pbaixo,pmedio,palto,pmalto}
@attribute torax {tbaixo,tmedio,talto,tmalto}
@attribute volt {vtbaixo,vtmedio,vtalto,vtmalto}
@attribute periodo {no-fatted,fatted}

@data
bmedio,pbaixo,talto,vtmedio,no-fatted
balto,palto,tmalto,vtmalto,fatted
bmalto,palto,talto,vtalto,fatted
bbaixo,palto,tmalto,vtalto,fatted
bbaixo,pmalto,tmalto,vtmedio,no-fatted
balto,palto,tmedio,vtalto,fatted
bbaixo,palto,tbaixo,vtmalto,no-fatted
balto,pbaixo,tmalto,vtbaixo,no-fatted
bmalto,pmalto,tmalto,vtbaixo,fatted
bmalto,pmalto,talto,vtmalto,fatted
balto,pmalto,tmalto,vtalto,fatted
bbaixo,pmalto,tmedio,vtmalto,no-fatted
bmedio,palto,tmedio,vtbaixo,no-fatted
bmedio,pbaixo,tbaixo,vtbaixo,no-fatted
balto,pmalto,tmedio,vtalto,fatted
bmedio,palto,tmalto,vtbaixo,no-fatted
balto,pmalto,talto,vtbaixo,fatted
```

Fonte: Arquivo pessoal, 2017

O processo de treino e testes são expressados pelos parâmetros de aspectos físicos e de qualidade seminal, os quais foram subdivididos em 3 subconjuntos de análises: o subconjunto 1 de aspectos físicos (braço, peso, tórax e volume testicular) e 2 subconjuntos de qualidade seminal, sendo eles: o subconjunto 2 de classificação morfológica (volume, grau de coagulação, motilidade, vigor, concentração e morfologia normal) e o subconjunto 3 morfologia normal (defeitos maiores, defeitos menores e morfologia normal). Outro ponto importante a ser mencionado é que a qualidade seminal tem sua própria categorização dos atributos já definida em literatura para cada métrica. Logo, foi utilizada a mesma para a codificação dos dados.

O primeiro treino foi realizado no subconjunto 1 que teve a finalidade de familiarizar os algoritmos com os seus períodos'' no-*fatted''* e *"fatted"* pré-estabelecidos. Assim, essa atividade organizou e treinou os dados gerados através de 15 relações distintas: 4 delas relacionam 1 atributo de maneira isolada ao período "*fatted*", 6 delas fazem a relação de 2 atributos ({Braço, Peso}, {Braço, Tórax}, {Braço, Volume testicular}, {Peso, Tórax}, {Peso, Volume testicular}, {Tórax, Volume testicular}) sem que haja repetições entre os atributos, 4 delas fazem relação de 3 atributos ({Braço, Peso, Tórax}, {Braço, Peso, Volume testicular}, {Braço, Tórax ,

Volume testicular}, {Peso, Tórax, Volume testicular}) de maneira distinta e a última leva em consideração todos atributos. Tudo isso foi feito com o objetivo de melhorar e dá uma maior confiabilidade aos resultados obtidos. Tanto para algoritmo *Apriori* quanto para algoritmo *FP-Growth.*

Vale frisar que os treinos e testes realizados especificamente no algoritmo *FP-Growth* necessitaram que duas novas transformações fossem feitas em todos os conjuntos de dados, tanto os de aspectos físicos, quanto para os de qualidade seminal sem exceções. Essa transformação se deu através da própria ferramenta WEKA, mediante aos algoritmos não supervisionados: *nominalToBinary* e o *numericToBinary*, os quais particionam ou detalham as variáveis de forma mais precisa.

Posteriormente, o subconjunto 2 de classificação morfológica é utilizado para o treinamento dos algoritmos, o qual é constituído por seis atributos. Porém, o grau de coagulação é desconsiderado para a etapa de identificação dos períodos reprodutivos. Porque não há nenhuma relação desse parâmetro com o período de reprodução, ou melhor, não há um grau de coagulação ideal. No entanto, seria de suma importância identificar qual ou quais graus de coagulação ocorrem de maneira mais frequentes no período "*fatted''* para os pesquisadores da área.

Por fim, é realizado o terceiro treino com o subconjunto 3 onde os dados são organizados em três atributos e seguem o mesmo requisito de treinamento do subconjunto 1 de aspectos físicos onde é feita a relação de 1 único atributo ao período "*fatted*", 2 atributos relacionados de forma distinta e assim sucessivamente até que todos atributos sejam usados.

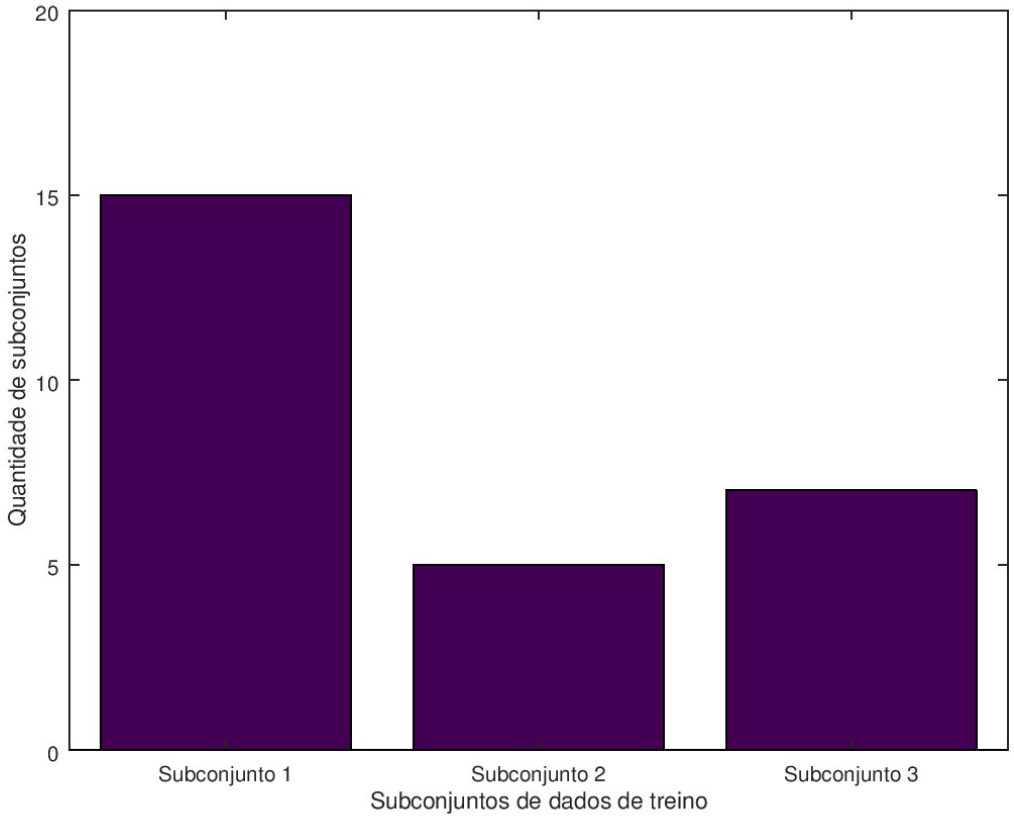
Outro ponto importante a ser frisado, é que os dados de morfologia além de terem suas próprias categorizações, foi preciso introduzir duas novas classificações nos dados, sendo elas: dados NEGATIVOS e a de dados POSITIVOS SEM (ou PSEM), ou seja, essa nomenclatura diz a respeito às coletas de sêmen que não foram bem sucedidas. Por exemplo: é realizado processo de coleta de sêmen, porém, não se houver a ejaculação esse acontecimento é referenciado como NEGATIVO para esta pesquisa bem como, aquelas coletas que a ejaculação acontece, no entanto, não há produção de nenhuma das propriedades seminais normalmente encontradas no sêmen, como: motilidade, vigor e concentração dentre outras, essa ocorrência é

definida como PSEM, já que houve a ejaculação. Porém, sem a presença das propriedades seminais.

4.5. DESCRIÇÃO DOS TESTES COM OS DADOS REAIS

Os testes nos dados reais foram feitos tanto no algoritmo *Apriori* quanto no *FP-Growth*, a etapa testes foi realizada depois que os algoritmos foram treinados. Assim, a regra de associação utilizou 27 conjuntos de dados de treino ao todo divididos da seguinte forma (Figura 8) 15 do subconjunto 1 de aspectos físicos, 5 do subconjunto de morfologia normal e 7 dos subconjuntos de classificação morfológica.

Figura 9. Conjunto de dados de treinamento



Fonte: Arquivo pessoal, 2017

Cada subconjunto de dados foi utilizado para treinar os algoritmos assim os algoritmos treinados eram submetidos a testes com os dados reais para detectar padrões e comprovar os paradigmas levantados pelos pesquisadores da área, ou seja, foi usado um conjunto de dados de treino e logo depois do treino ocorria os testes

com os dados reais. Esse processo de identificação buscava determinar qual era a estação reprodutiva dos macacos da espécie *Saimiri Collinsi* que vivem em cativeiro, por meio dos parâmetros de aspecto físico.

Vale mencionar que o mesmo processo de treinamento e os testes de dados foi realizado nos subconjuntos de dados de qualidade seminal para que também pode-se identificar o período reprodutivo dos animais e o melhor período para a coleta do sêmen. Essa etapa foi feita através de 12 conjuntos de dados dividas em 2 subconjuntos.

4.6. RESULTADOS DOS PARÂMENTROS DE ASPECTOS FÍSICOS

Em todas as análises de mineração de dados foi utilizada a regra de associação, tanto para os dados de aspecto físico, quanto para os dados de qualidade seminal, pois as regras de associação são capazes de detectar a relação entre itens em uma base de dados. Assim a utilização de uma única técnica foi em decorrência da problemática deste trabalho que visa relacionar quais parâmetros de aspecto físicos estão correlacionados a período reprodutivo (ou ”*fatted”*) dos *Saimiris Collinsi,* que vivem em cativeiro. Com isso, os treinos e a mineração propriamente ditam foi realizada através dos algoritmos *Apriori e FP-Growth,* conforme foi mencionado anteriormente.

Consequentemente, foram geradas 53 regras pelo algoritmo *Apriori* referente aos subconjuntos de aspectos físicos ao todo onde o grau de confiança das regras varia entre 93% a 60% demonstrando quais os atributos com maior relação de incidência ao período “*fatted*”. Assim, foi possível analisar e avaliar cada regra gerada junto aos especialistas da área. Vale resultar que as análises deste trabalho ocorreram da seguinte forma: cada resultado obtido foi comparado e avaliado aos resultados estatísticos realizados em paralelo ao deste trabalho pelos pesquisados da área.

Tabela 4. Regras encontradas pelo algoritmo *Apriori*

Regras	Braço	Peso	Tórax	Vol. Test.	Resultado	Confiança
1	BMALTO	SEM	BMALTO	SEM	“FATTED”	93%
2	BALTO	PALTO	SEM	SEM	“FATTED”	91%
10	SEM	PALTO	TALTO	SEM	“FATTED”	79%
11	BALTO	SEM	SEM	VTALTO	“FATTED”	78%
12	SEM	PALTO	SEM	VTALTO	“FATTED”	76%
14	BALTO	SEM	ALTO	SEM	“FATTED”	74%
17	SEM	PALTO	SEM	SEM	“FATTED”	73%
23	BMALTO	SEM	TMALTO	SEM	“FATTED”	70%
35	SEM	SEM	SEM	VTMALTO	“FATTED”	67%
36	BALTO	PALTO	SEM	SEM	“FATTED”	67%
37	BALTO	SEM	TALTO	SEM	“FATTED”	67%
39	SEM	SEM	SEM	ALTO	“FATTED”	66%
38	BMALTO	SEM	TMALTO	SEM	“FATTED”	67%
42	BALTO	SEM	TALTO	SEM	“FATTED”	65%
43	BALTO	SEM	TALTO	SEM	“FATTED”	65%
52	SEM	SEM	TALTO	SEM	“FATTED”	60%

Legenda: B: BRACO; P:PESO; T: TÓRAX; VT: VOLUME TESTICULAR; MA: MUITO ALTO

Fonte: Arquivo pessoal, 2017

É possível presumir que a partir da regra 1 Tabela 3, com o grau de confiança de 0.93, que a relação entre os atributos dobra cutânea do braço e tórax categorizados em medidas proporcionalmente muitas altas têm influência direta no período reprodutivo. Dessa forma, o aumento das dobras cutâneas desses parâmetros estabelece que o animal está em período “*fatted”*.

As regras 2, 10, 12, 17 e 36 Tabela 3 confirmaram um dos paradigmas levantados no processo de catalogação dos dados que estabelece que se as medidas do peso forem altas e/ou muito altas de maneira isolada ou relacionada a outras medidas de outros parâmetros com os intervalos proporcionalmente iguais ao seu,

pode-se assumir que os macacos-de-cheiro estão em período reprodutivos com o grau de confiança entre 91% a 67% do percentual de certeza. Essa métrica tem a maior ocorrência nos casos onde os períodos são classificados com "*fatted"*.

Outro padrão interessante, está associado aos conjuntos de regras {2, 11, 14, 36, 37, 42, 43} e {1, 23, 38} Tabela 3, os quais são responsáveis por determinar que dobras cutâneas do braço categorizadas como: altas e muito altas, sejam fatores essenciais para indicar que os animais estão em período reprodutivos "*fatted*", ou seja, as medidas que encontrarem-se em intervalos previamente estabelecidos como: alto e muito alto respectivamente estando associadas a outros atributos podem ser um indicativo que os macacos-de-cheiro encontram-se em período de reprodutivo essa métrica também pode ser entendida como a segunda métrica de maior importância na definição do período reprodutivo, pois de maneira isolada ela não consegue definir o período "*fatted"*, no entanto, na maioria das regras encontradas como o período '' *fatted* '' ela está presente em boa parte.

Por fim, temos as métricas da dobra cutânea tórax e volume testicular as quais estabelecem relações com os parâmetros da dobra cutânea braço e peso com medidas respectivamente altas e muito altas. Assim, podemos entendê-las como métricas secundárias nesse processo de identificação de padrões. Pois, de forma isolada nenhuma das regras {52, 35, 39} tem confiabilidade maior 0.70 por cento do grau de confiança em seus resultados.

Sob outra perspectiva de análise realizada nos conjuntos de dados de treino e nos conjuntos de dados reais, realizada no algoritmo *FP-Growth* da ferramenta WEKA, a qual tenta identificar padrões de maneira mais precisa em uma base de dados transacional, onde suas análises inicialmente processam a identificação dos itens com menor grau de frequência, pois esses itens são removidos do processo de descoberta de conhecimento para um melhor desempenho algorítmico, ou seja, o conjunto de dados é diminuído e assim os demais itens são organizados de forma hierárquica do mais frequente até o menos frequente. Com isso os resultados encontrados no *algoritmo PF-Growth* variam entre 93% a 61% o seu grau de confiança das regras obtidas ao todo foram geradas 36 regras que define o período "*fatted".*

Tabela 5 - Regras encontradas pelo algoritmo *FP-Growth*

Regras	Braço	Peso	Tórax	Vol. Test.	Resultado	Confiança
1	BMALTO	SEM	TMALTO	SEM	“FATTED”	93%
2	BALTP	PALTO	SEM	SEM	“FATTED”	91%
8	SEM	PALTO	TALTO	SEM	“FATTED”	79%
10	SEM	PALTO	SEM	VTALTO	“FATTED”	76%
12	SEM	PALTO	SEM	SEM	“FATTED”	73%
15	BMALTO	SEM	TMALTO	SEM	“FATTED”	70%
22	SEM	SEM	SEM	VTMALTO	“FATTED”	67%
25	BALTO	ALTO	SEM	SEM	“FATTED”	67%
26	BMALTO	SEM	TMALTO	SEM	“FATTED”	67%
27	SEM	SEM	SEM	VTALTO	“FATTED”	66%
30	BALTO	PALTO	SEM	SEM	“FATTED”	65%

Legenda: B: BRACO; P:PESO; T: TÓRAX; VT: VOLUME TESTICULAR; MA: MUITO ALTO

Fonte: Arquivo pessoal, 2017

As regras 1, 15 e 26 (Tabela 4) estabelecem que os atributos: dobra cutânea do braço e tórax que são classificados com medidas muito altas pode-se assumir que os macacos-de-cheiro se encontra em período reprodutivo, como o grau de confiança entre 0.93 a 0.67 através das análises do *Algoritmo FP-Growth,* como também foi identificado pelo algoritmo *Apriori*.

As regras 2, 8, 10, 12, 25 e 30 (Tabela 4), define que os atributos do peso categorizados com medidas altas e muito altas, podem definir que macacos-de-cheiro estão no período “*fatted”* estando ela isolada ou relacionada a outros parâmetros com a mesma nomenclatura de categorização aplicada a ele. Vale ressaltar que o mesmo resultado foi encontrado para os testes como reais com o grau de confiança 91% a 61% nas regras.

As regras 22 e 27 (Tabela 4) estabelecem que medidas do volume testicular, consideradas altas e muito altas, podem determinar o período “*fatted”* dos animais de forma isolada desde que esse parâmetro esteja de acordo com essas medidas descritas acima. Por fim, as regras que associam os parâmetros do braço com o

período reprodutivo dos animais estudados expressam-se da seguinte forma: as métricas forem categorizadas como altas e muito altas podem ser indício que o animal está em período" *fatted*".

4.7. RESULTADOS OBTIDOS DA QUALIDADE SEMINAL

Assim, como a primeira atividade de identificações de padrões realizada sobre o conjunto de dados referentes ao subconjunto de aspectos físicos através dos algoritmos *Apriori* e *FP-growth* os parâmetros da qualidade seminal também passaram por duas análises algorítmicas. No entanto, o algoritmo *FP-growth* obteve um melhor desempenho no processo de identificação de padrões, porque seus resultados foram semelhantes aos resultados encontrados pelos pesquisadores da área que submeteram os parâmetros de qualidade seminal a análises estatísticas com visões diferentes as aplicadas a essa pesquisa, os quais diferiam entre os períodos com o peso menor ($p < 0.05$) entre o ("*no-fatting*" e "*fatting*") para as variáveis: pH, motilidade, integridade da membrana plasmática e defeitos maiores em sua analises.

Vale ressaltar, que tanto o subconjunto 2, quanto o subconjunto 3 geraram regras que indicam que independentemente do animal está ou não em seu período reprodutivo a coleta de sêmen poderá ser feita, porque a estação reprodutiva parece não interferir nas espermatogêneses nem na qualidade seminal dos animais, sendo essa uma análise feita pelos pesquisadores da área em ambas as pesquisas.

Outro ponto importante a ser frisado, referente às atividades de coleta do sêmen, para aquelas amostras que não apresentam a ejaculação as regras sugerem que os macacos-de-cheiro estão em período "*fatted"* mesmo sem ejacular, pois esse evento está associado, ao fato dos os primatas não-humanos serem conhecidos por se masturbarem em condições cativas (HARRISON, 1980), podendo afetar diretamente no processo de coleta seminal e da concentração espermática.

Essa afirmação constata o seguinte paradigma evidenciado pelos pesquisadores da área, o qual sugere que os macacos-de-cheiro na estação reprodutiva por se trata de um gênero com sazonalidade bem definida como foi mencionado na seção 2 e copulando mais vezes que o normal na estação de acasalamento isso faz com que a produção do sêmen seja normalmente diminuída da concentração espermática devido a esse acontecimento.

SEÇÃO 5

CONCLUÇÕES E TRABALHOS FUTUROS

Nesta seção, são apresentadas as considerações finais e as dificuldades encontradas para realização deste trabalho. Consequentemente, são sugeridos alguns trabalhos futuros que podem ser realizados nesta área de conhecimento.

5.1. CONSIDERAÇÕES FINAIS

Neste trabalho foi aplicado o processo de descoberta de conhecimento em bases dados KDD em um problema real, da Medicina Veterinária utilizando a etapa de mineração de dados para a identificação e descoberta de padrões. Que tentou descobrir e constatar quais características físicas e morfológicas estavam correlacionadas diretamente com o período reprodutivo dos macacos-de-cheiro do da espécie *Saimiri Collinsi*, que vivem em cativeiro.

Logo após, a essa distinção entre os períodos não-reprodutivos e reprodutivos, novas análises foram feitas sob os dados coletados, no entanto, essa nova análise buscou identificar quais parâmetros da qualidade seminal, sofriam influência direta em suas propriedades correlacionadas ao período *''fatted''*. Porém, devido às mudanças comportamentais apresentadas pelos animais desta pesquisa, algumas amostras do sêmen não apresentavam algumas das suas propriedades seminais.

Assim as análises e as avalições dos resultados obtidos foram submetidas uma comparação entre os resultados dos algoritmos, onde o algoritmo *FP-Growth* teve um melhor desempenho frente ao algoritmo *Apriori* de 90% pois os seus resultados são semelhantes aos resultados encontrados pelos pesquisadores de Medicina Veterinária através de métodos de descoberta de padrões estatísticos realizados paralelamente aos desses trabalhos. Além de apresenta os melhores resultados o *FP-Growth* aplicar uma remoção 50% dos conjuntos de dados para este problema assim melhorando o desempenho computacional.

No decorrer desta pesquisa alguns problemas foram encontrados, como por exemplo: a pouca quantidade de dados reais (DR) que foi solucionado através da

geração de dados pseudo-aleatórios (DS) através da distribuição normal de *Gauss.* Outro ponto importante a ser solucionado seria relacionar os parâmetros de qualidade seminal com o *''fatted'',* pois a quantidade de atributos classificados como *''fatted''* para essa análise era pequena e poderia influenciar diretamente nos resultados obtidos de forma negativa.

Então, esta etapa de seleção dos dados tornou-se longa e criteriosa com o constante acompanhamento dos especialistas da área. Assim, garantiu que todo o processo de análise, bem como a seleção, preparação, transformação dos dados fosse bem sucedida até que a mineração de dados dita pudesse ser aplicada, pois a retirada de qualquer dado de forma errada poderia retornar resultados equivocados.

Em suma, a partir deste trabalho de exploração de dados com a finalidade de identificar informações úteis que serviram para estudos futuros nota-se que esse processo de busca é custoso e demanda bastante tempo e controle nas atividades, as quais devem ser realizadas de forma simples. Porém, a etapa de mineração de dados apresenta algumas complexidades em sua aplicação que estão relacionadas com a preparação dos dados como um todo.

Nesse sentido, o processo de preparação dos dados consistiu na etapa de geração dos dados sintéticos (DS), na concepção da melhor técnica de geração de dados sempre levando em conta a confiabilidade dos resultados, assim o processo de preparação dos dados foi a etapa que demandou uma quantidade expressiva de tempo, tanto de quem aplicava as técnicas de mineração, quanto dos especialistas da área que avaliavam as regras encontradas, assim este trabalho e considerado válido para solucionar a problemática em questão.

5.2. TRABALHOS FUTUROS

Levando em consideração que a coleta e o armazenamento dos dados desta pesquisa de forma semi- automatizada em planilhas da ferramenta Excel, uma alternativa simples, porém eficiente seria o desenvolvimento de uma aplicação móvel, a qual estaria conectada diretamente a um banco de bandos distribuído que pudesse ser acessado a qualquer momento, caso necessário, facilitando futuras análises de extração de padrões e dando uma maior segurança para os dados.

Assim, com a implementação dessa aplicação móvel juntamente ao um banco de dados distribuído, os pesquisadores de Medicina Veterinária, teriam maior controle no manuseio os dados catalogados em suas pesquisas. Outro ponto importante seria especificamente fazer uso de técnicas automatizadas para extração de padrões a partir de imagens do sêmen dos macacos sem precisar necessariamente de analises manuais.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. *Mining Association Rules Between Serts of Itens in Large DataBases.* ACM SIGMOD conference Management of Data. 1993.

ALMEIDA, F.C; DUMONTIE ouR, P**.** O Uso De Redes Neurais Em Avaliação De Riscos De Inadimplência, Revista de Administração FEA/USP, São Paulo, jan/mar 1996.

ANDRADE, A. F. C. et al. Fluorescent stain method for the simultaneous determination of mitochondrial potential and integrity of plasma and acrosomal membranes in boar sperm. **Reproduction in Domestic Animals**, v. 42, n. 2, p. 190-194, 2007.

ANDRADE, A. F. C.; ARRUDA, R. P.; CELEGHINI, E. C. C.; NASCIMENTO, J.; MARTINS, S. M. M. K.; RAPHAEL, C. F.; MORETTI, A. S. Fluorescent Stain Method for the Simultaneous Determination of Mitochondrial Potential and Integrity of Plasma and Acrosomal Membranes in Boar Sperm. Reproduction in Domestic Animals, v. 42, p. 190-194, 2007.

AURICCHIO, P. **Primatas do Brasil**. Sao Paulo:Terra Brasilis, 168 p, 1995.

BALDWIN, J. D. *The Behavior of Squirrel Monkeys (Saimiri) in Natural Environments*. In: ROSENBLUM, L. A.; COE, C. L. Handbook of Squirrel Monkey Research. New York: Plenum Press, p. 35-53, 1985.

BLOM, E. The ultrastructure of some characteristic sperm defects and a proposal for a new classification of the bull spermogram. **Nordisk Veterinaermedicin**, v. 25, n.7, p. 383-391, 1973.

CASSIA, de cássia Davi das Neves, Pré-Processamento De Descoberta De Em Banco De Dados, 2003. Dissertação (Mestrado em Ciência da Computação), Universidade Federal do Rio Grande do Sul. 2003.

COLÉGIO BRASILEIRO DE REPRODUÇÃO ANIMAL (CBRA). **Manual para exame andrológico e avaliação de sêmen animal**. 3ª Ed. Belo Horizonte, 2013. p. 49.

DIXSON, A. F.; ANDERSON, M. J. *Sexual selectin and the comparative anatomy of reproduction in monkeys, apes, and human being. Annual Review of SexResearch*, v.12, p. 121-144, 2001.

DUKELOW, W. R. *The Squirrel Monkey (Saimiri sciureus). In: HEARN, J. P. Reproduction in New World Primates: New Models in Medical Science.Lancaster: MTP Press Limited,* 1983, p. 149-180.

DUMOND, F. V. and HUTCHINSON, T. C. *Squirrel monkey reproduction: the “’fatted’” male phenomenon and seasonal spermatogenesis. Science*, v. 158, p. 1067-1070, 1967.

Eibe frank; Mark a. Hall; Ian H. Witten. *The WEKA Workbench “Data Mining: Proctical machine learning Tools and Techniques”. Morgan Kaufmann, Fourth Edition*, 2016.

ELMASRI, Ramez; NAVATHE, Shamkant. Sistemas de Banco de Dados. 4ª Ed. Sao Paulo: *Addison-wesley*, 2005, p. 625- 637.

FALEIRO; Christiane Sidney. Aplicação De Mineração De Dados No Banco De Dados Do Zoneamento Ecológico Econômico De Minas Gerais, 2010. Monografia (Bacharelado em Sistemas de Informação), 2010.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P; UTHURUSAMY, R.. *Advances in Knowledge Discovery and Data Mining. USA*, California: AAAI Press / MIT Press, 1996.

FRANCIELLE, Aline dos Anjos Lima, Descoberta de Padrões no Sêmen de Primatas Não-Humanos Através da Mineração de Dados. 2014. Teste Conclusão de Curso (Bacharelado em Sistemas de Informação), Universidade Federal do Pará. 2014.

Galo, M.; Camargo, P. D. Oliveira. Tutorial do Octave / OCTAVE Tutorial. Universidade estadual Paulista, Departamento de Cartografia, Presidente Prudente, 2016.

GARNER, D. L. Flow cytometric sexing of mammalian sperm. **Theriogenology**, v. 65, n. 5, p. 943-957, 2006.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA**,** Eduardo. **Data Mining:** conceito, técnicas, algoritmos orientações e aplicações. 2. ed. Rio de janeiro: Elsevier, p. 1- 4- 22- 23- 30, 2015.

GOMES, Luciana Castanheira. Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação De Padrões**.** 2008. Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal Minas Gerais. 2008.

Han, J. and Kamber, M. (2006) Data mining: concepts and techniques. San Francisco, Morgan Kaufmann.

HAND, D; MANNILA, H; SMYTH, P. *Principles of Data Mining*. *MIT Press*, 2001.

Harrison, R. 1980. Semen parameters in *Macaca mulatta:* ejaculates from random and selected monkeys. ***J. Med. Primatol****.,* 9:265 - 273.

HAYKIN, S. Neural Networks: *a interactive Data Analysis: The Control Project*. IEEE Transctions on Computer Socitey, v. 18, 1999.

ICMBio, Instituto Chico Mendes de Conservação de biodiversidade. Avaliação do Risco de extinção de Saimiri sciureus (Linnaeus, 1758). Disponível em:<

http://www.icmbio.gov.br/portal/faunabrasileira/estado-de conservacao/7266-mamiferos-saimiri-sciureus-macaco-de-cheiro>. Acessado em: 22/04/2017.

INGBERMANN, B.; STONE, A.I.; CHEIDA, C.C. Gênero Saimiri (VOIGT, 1831). In: REVIS, N. R.; PEROECH, A.L; ANDRADE, F.R. (orgs). Primatas do Brasil. Londrina:Technical Book Editora, 2008. p. 41-46.

KUGELMEIER, T.; VALLE, R. R.; MONTEIRO, F. O. B. Biologia da reproducao. In:ANDRADE, A.; ANDRADE, M. C. R.; MARINHO, A. M.; FILHO, J. F. Biologia, Manejo e Medicina de Primatas Não Humanos na Pesquisa Biomédica. Rio de Janeiro: Editora Fiocruz, p. 68, 2010.

KUGELMEIER, Tatiana. Colheita e análise do sêmen de macacos de cheiro (*Saimiri sciureus*) por vibroestimulação: do condicionamento ao coágulo seminal, Tese (doutorado em reprodução animal). Universidade de São Paulo. Faculdade de Medicina Veterinária e Zootecnia, Departamento de Reprodução Animal, São Paulo, 2011.

MOCE, E.; GRAHAM, J. K. In vitro evaluation of sperm quality. **Animal Reproduction Science**, v.105, n. 1, p. 104-118, 2008.

OLIVEIRA, Cassio Camilo; CARLOS, João da Silva. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas, 2009. Relatório Técnico (Mestrado em Ciência da Computação), Instituto de Informática Universidade Federal de Goiás. 2009.

PAULINHO, G. L. Estudo anatômico-histológico descritivo do sistema reprodutor feminino de três espécies do gênero *Saimiri* Voigt, 1831 (Primates, Cebidae). Dissertação (mestrado em ciência animal), Universidade Federal do Pará. 2013.

ROSENBLUM, L.A. *Some aspects of female reproductive physiology in the squirrel monkey*. In: ROSENBLUM, L.A; *COOPER, R. W. The Squirrel Monkey. New York: Academic Press*, 1968. p.147-169.

SENGER, P.L. Reproductive cyclicity – terminology an basic concepts. In: Pathways to pregnancy and parturition. 2ed. Current conceptions p. 144-163, 2003.

Silva, Leandro Augusto da. Introdução à mineração de dados: com aplicações em R / Leandro Augusto de Silva, Serajane Marque Peres, Clodis Boscarrioli. – 1. ed. – Rio de Janeiro: Elsevier, 2016.

STEINBERG, E. R.; PALERMO, A. M.; NIEVES, M.; BURNA, A.; SOLIS, G.; ZUNINO, G.; MUDRY, M. D. *Sex determination and sperm morphology in cebidae*. Anais do XI Congresso brasileiro de primatologia. Porto Alegre, 2005.

STONE, A. I. Responses of squirrel monkeys to seasonal changes in food availability in an Eastern Amazonian rainforest. American Journal of Primatology, v. 69, p. 142- 157, 2007.

TAN, Pang-Ning; STREINBACH, Michael; KUMAR, Vipin. Introdução ao DATA MINING Mineração de dados, Rio de Janeiro: Editora Ciência Moderna Ltda, p.4- 6- 49- 270-392-409-874, 2009.

VIANA, CLÊNIO FERNANDES. Características do sêmen, perfil da concentração de testosterona no extrato fecal, variação da massa corporal e volume testicular de micos-de-cheiro (*saimiri sciureus,* linnaeus*,* 1758*)* mantidos em cativeiro sob condições ambientais controladas. Dissertação (Mestrado em ciência animal). Universidade estadual do norte Fluminense Darcy Ribeiro, Centro de Ciências e Tecnologias Agropecuária. Campos dos Goytacazes, RJ, 2013.

WILLIMS, L. Aging Cebidae In: ATSALIS, S.; S.W. MAGULIS, S.W. *Primate Reproductive Aging: cross-taxon perspectives*. Switzerland: Ed Karger v. 36, p.49-61, 2008.

YEOMAN, R. R.; SONKSEN, J.; GIBSON, S. V.; RISK, B. M.; ABEE, C. R. Penile vibratory stimulation yields increased spermatozoa and accessory gland production compared with rectal electroejaculation in a neurologically intact primate (Saimiri boliviensis). Human Reproduction, v. 13, n. 9, p. 2527-2531, 1998.

ZAMBELLI, D.; CUNTO, M. Semen collection in cats: Techniques and analysis. **Theriogenology**, v. 66, n. 2, p. 159-165, 2006.