UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

ANTONIO RAFAEL BRAGA

MODELOS DE CLASSIFICAÇÃO PARA PREDIÇÃO DO BEM ESTAR DE COLÔNIAS DA ABELHA *APIS MELLIFERA*

FORTALEZA

2020

ANTONIO RAFAEL BRAGA

MODELOS DE CLASSIFICAÇÃO PARA PREDIÇÃO DO BEM ESTAR DE COLÔNIAS DA ABELHA *APIS MELLIFERA*

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Danielo Gonçalves Gomes

FORTALEZA

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B792m Braga, Antonio Rafael.
Modelos de Classificação para Predição do Bem Estar de Colônias da Abelha Apis mellifera / Antonio Rafael Braga. – 2019.
124 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2019.
Orientação: Prof. Dr. Danielo Gonçalves Gomes.

1. Ciência de dados. 2. Mineração de dados. 3. Apicultura de precisão. 4. Bem estar de abelhas. 5. Apis mellifera. I. Título.

CDD 621.38

ANTONIO RAFAEL BRAGA

MODELOS DE CLASSIFICAÇÃO PARA PREDIÇÃO DO BEM ESTAR DE COLÔNIAS DA ABELHA *APIS MELLIFERA*

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 02 de Março de 2020

BANCA EXAMINADORA

Prof. Dr. Danielo Gonçalves Gomes (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Michela Mulas
Universidade Federal do Ceará (UFC)

Profa. Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. Breno Magalhães Freitas
Universidade Federal do Ceará (UFC)

Prof. Dr. Gustavo Pessin
Instituto Tecnológico Vale (ITV)

Prof. Dr. Walmes Marques Zeviani
Universidade Federal do Paraná (UFPR)

Dedico esse trabalho a Deus.

AGRADECIMENTOS

Gratidão a Deus, pela minha vida, saúde e paz, e por criar as abelhas.

Gratidão aos meus pais (Lúcia e Braga), às minhas irmãs (Mirla e Mirlane), às minha sobrinhas (Fernanda e Julia) pela confiança, cuidado e amor que demostram por mim. Gratidão à Jerí, ao André, ao Fernando, ao Thiago, a Lívia, a Gizele, ao Elano e a D. Silvia pelo apoio.

Agradeço ao Prof. Dr. Danielo G. Gomes, pela amizade, pelo apoio nos momentos difíceis, por me conduzir no mestrado, por me aceitar como aluno no doutorado. Enfim, por ter me orientado nessa caminhada. Obrigado pela confiança em mim depositada.

Agradeço aos servidores docentes e técnicos administrativos do campus da UFC em Quixadá, em especial, ao Prof. Dr. Davi Romero de Vasconcelos e a Profa. Dra. Andreia Liborio Sampaio, que, enquanto gestores, me concederam afastamento das minhas atividades de ensino para realização essa trabalho.

Agradeço à coordenação, a todos os professores e à secretaria do PPGETI pelo apoio ao longo de todo o doutorado e pela contribuíção direta ou indireta na formação acadêmica.

Agradeço à Profa. Dra. Rossana M. C. Andrade, coordenadora do laboratório GREat, por me disponibilizar uma moderna infraestrutura laboratorial. Agradeço também aos diversos membros do GREat que interagi ao longo desses quatro anos, em especial, ao: Valdenir, Breno, Belmondo, Tales, Italo, Rodrigo, Joseane, Paulo Arthur, Atslands, Darilu, Jonas, Rute, Bruno Sabóia, Jhonson, Fábio, André, Vandebergue, Nídia e Manuel.

Agradeço a todos que fizeram ou fazem parte do projeto Sm@rtBee, muito obrigado pelo apoio, e, nominalmente, ao: Felipe, Samuel, Leandro, Sara, Gustavo, Rhaniel, Lia, Lucas Noleto, Alisson, Juliana, Alef, Lucas Esteves, Cid, Davyd, Ícaro, Daniel, Aluísio, Júnior, Felipe e Leo. Agradeço ao Prof Dr. Breno M. Freitas e ao Dr. Diego Bezerra por todo apoio e orientação em relação aos temas relacionados às abelhas. Ao CNPq pelo apoio financeiro ao projeto.

Agradeço ao Prof. PhD Joseph A. Cazier, Prof. PhD Ed Hassler e ao Prof. MSc Dick Rogers por todo apoio e orientação no meu período de estagio na Appalachian State University. Obrigado também ao Andrew J. Scott, Preston MacDonald e ao Josh Williams. Thank you!

Agradeço aos professores PhD Breno Magalhães Freitas, PhD Michela Mulas, Dra. Ticiana Linhares Coelho da Silva, Dr. Gustavo Pessin e Dr. Walmes Marques Zeviani por participar da banca examinadora dessa tese e pelas preciosas contribuições.

”Como eu vos amei, vocês devem amar-se uns aos outro”

(João Evangelista, 13:34, *apud* Jesus Cristo)

RESUMO

As abelhas são os principais polinizadores das espécies de plantas silvestres polinizadas por insetos e são essenciais para a manutenção dos ecossistemas vegetais e para a produção de alimentos. No entanto, nas últimas três décadas, elas sofreram inúmeros desafios com relação ao bem estar, incluindo mudanças climáticas, poluentes, toxinas, pragas e doenças, em especial a espécia *Apis mellifera*, uma das mais importantes espécies de abelhas para a polinizaçao. Uma tentativa de mitigar esse problema é estimar o nível de bem estar das colônias e indicar um estado de colapso iminente para os apicultores. Para estimar o nível de bem estar das colônias de abelhas, aplicamos três métodos de análise de dados que calibram algoritmos de classificação e regressão com base em abordagens de aprendizado de máquina supervisionado e não supervisionado. Para validar o primeiro método aplicado, foi utilizado um conjunto de dados reais de duas colmeias obtidas no portal HiveTool.net com temperatura da cria, umidade relativa e peso de colmeias de abelhas *Apis mellifera*. A partir do índice *Calinski-Harabasz* e do algoritmo *k-means*, foram encontrados 6 padrões sazonais de colônia relacionados às transições entre as estações do ano. A partir dos padrões encontrados, três algoritmos de classificação foram treinados, validados e testados. Para validar o segundo método, foi usado um conjunto de dados obtido de 6 apiários, totalizando 27 colmeias de abelhas *Apis mellifera* monitoradas ao longo de três anos. Três algoritmos de classificação foram treinados, validados e testados. Em termos de atributos, foi utilizada a temperatura interna e o peso da colméia, além de dados climáticos (temperatura, ponto de orvalho, direção do vento, velocidade do vento, precipitação e luz do dia). Também foram usadas 703 inspeções *in loco* de apiário feitas semanalmente para adicionar aos dados dos sensores o rótulo necessário na fase de treinamento dos algoritmos de classificação. Finalmente, para validar o terceiro método, o algoritmo *Long Short-Term Memory* (LSTM) foi aplicado um conjuntos de dados reais obtido através do sistema de monitoramento remoto de colméias Arnia. O conjunto de dados possui dados de temperatura da cria (temperatura interna), umidade interna, ventilação média, ruído médio de vôo, peso da colméia e temperatura externa coletada ao longo do outono europeu em 2017. Os resultados obtidos com a aplicação dos métodos sugerem que os algoritmos de classificação e regressão são eficientes para a obtenção de modelos de alta precisão para predição de níveis de bem estar das colônias de abelhas *Apis mellifera*.

Palavras-chave: Apicultura de precisão. *Apis mellifera*. Ciência de dados. Clusterização. Classificação. Bem estar. Abelhas.

ABSTRACT

Bees are the main pollinators of most species of wild plants pollinated by insects and are essential for the maintenance of plant ecosystems and food production. However, in the past three decades, they have suffered numerous health challenges, including changes in habitat, pollutants, toxins, pests, diseases, and competition for resources. An attempt to mitigate this problem is to estimate the health status of the colonies and indicate a state of imminent collapse for beekeepers. To estimate the health status of bee colonies, we propose three methods of data analysis that calibrate classification and regression algorithms based on supervised and unsupervised machine learning approaches. To validate the first proposed method, a real dataset from two hives obtained from the HiveTool.net portal was used with internal temperature, relative humidity and weight of *Apis mellifera* beehives. From *Calinski-Harabasz* index and the *k-means* algorithm, 6 colony health patterns related to transitions between seasons were found. From the found patterns, three classification algorithms were trained, validated and tested. To validate the second method, a data-set obtained from 6 apiaries was used. In this data-set, 27 *Apis mellifera* beehives were monitored over three years. Three classification algorithms were trained, validated and tested. In terms of attributes, the internal temperature and the weight of the hive were used, in addition to climatic data (external temperature, dew point, wind direction, wind speed, precipitation, and daylight). Also, 703 *in loco* apiary inspections carried out weekly were also used to put labels in sensors data. Finally, to validate the third method, the *Long Short-Term Memory* (LSTM) algorithm was applied to a real data-set obtained through the Arnia remote monitoring system. The data-set has data on brood temperature (internal temperature), internal humidity, average ventilation, average flight noise, hive weight and external temperature collected throughout the European autumn in 2017. The results obtained with the application of the methods suggest that the classification and regression algorithms are efficient to obtain high precision models for predicting colony health levels.

Keywords: Precision beekeeping. *Apis mellifera*. Data mining. Clustering. Classification. Regression. Health levels. Bee colonies.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

ARIMA	*Auto-Regressive Integrated Moving Average*
AUC ROC	*Area Under Receiver Operating Characteristic Curve*
BBCC	*Bayer Bee Care Center*
BIP	*Bee Informed Partnership*
CAPA	*Canadian Association of Professional Apiculturists*
CCD	*Colony Collapse Disorder*
CH	*Calinski-Harabasz*
COLOSS	*Prevention of honey bee COlony LOSSes Association*
CV	*Cross-Validation*
EDA	*Exploratory Data Analysis*
HCC	*Healthy Colony Checklist*
HS	*Health Status*
IoT	*Internet of Things*
KDD	*Knowledge Discovery in Databases*
KDE	*Kernel Density Estimation*
kNN	*k-Nearest Neighbor*
L-BFGS	*Broyden-Fletcher-Goldfarb-Shanno Limited-memory*
LOF	*Local Outlier Factor*
LSTM	*Long Short-Term Memory*
MLP	*Multi Layer Perceptron*
MSE	*Mean Square Error*
NB	*Naive Bayes*
NN	*Neural Networks*
NWS	*National Weather Service*
RBF	*Radial Basis Feature*
ReLU	*Rectified Linear Unit*
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
RNN	*Recurrent Neural Network*
SSD	*Sum of Squared Differences*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

Esta tese aplica três métodos para mineração de conjuntos de dados de colmeias de abelhas. Os conjuntos possuem dados de inspeção, de sensores internos e sensores externos às colmeias e são usados para prever o nível de bem estar das colônias de abelhas. Os métodos são construídos e definidos com base em três das principais técnicas de mineração de dados, a saber: agrupamento, classificação e regressão. Os métodos são independentes em si, logo, cada método pode ser aplicado isoladamente e de acordo com o conjunto de dados selecionado. O primeiro método é baseado em agrupamento para mineração de padrões sazonais e é recomendado para conjunto de dados não rotulados. O segundo método é baseado na técnica de classificação e foi desenvolvido para classificar dados combinados de inspeções, de sensores internos e externos para prever o nível de bem estar das colônias de abelhas. O terceiro método é baseado em regressão, onde dados de sensores internos e externos às colméias são utilizados para prever perda de capacidade de termorregulação em colônias de abelhas. A contextualização, caracterização do problema, questões de pesquisa, hipóteses, objetivos, contribuições e estrutura do documento são descritos abaixo nesta introdução.

1.1 Contextualização

Entre todos os agentes polinizadores animais, a polinização realizada pelos insetos foi avaliada em 153 bilhões de euros em culturas de alimentos em todo o mundo, representando 9,5% do valor total da produção agrícola mundial usada para alimentação humana (GALLAI *et al.*, 2009; POTTS *et al.*, 2016). Estima-se que aproximadamente 75% das culturas em todo o mundo dependem dos insetos para a produção agrícola de frutas e/ou sementes (KEVAN; PHILLIPS, 2001; KLEIN *et al.*, 2007; OLLERTON *et al.*, 2011; POTTS *et al.*, 2016). Dentre todos os insetos polinizadores, as abelhas são consideradas os mais significativos, sendo possível observar um crescimento na dependência das abelhas para polinização de cerca de 300% nos últimos 60 anos (AIZEN; HARDER, 2009).

Além da produção agrícola, polinizadores animais, especialmente os insetos, são importantes agentes bióticos que, com o serviço de polinização, contribuem para a estabilidade dos ecossistemas, manutenção da biodiversidade de espécies de plantas silvestres (POTTS *et al.*, 2016; OLLERTON, 2017; BEZERRA *et al.*, 2019). No entanto, a diversidade de polinizadores está em declínio. Nas paisagens agrícolas, isso geralmente é observado em grandes monoculturas,

mas existe a preocupação de que a expansão urbana e o desmatamento também tenham um impacto negativo na diversidade de polinizadores (VERBOVEN *et al.*, 2014; LEVÉ *et al.*, 2019; SPONSLER *et al.*, 2019). Reduções na diversidade e abundância de polinizadores podem afetar a reprodução de espécies vegetais, produção agrícola, segurança alimentar e bem-estar humano (POTTS *et al.*, 2010; GARIBALDI *et al.*, 2013). Assim, na qualidade de principal agente polinizador, as abelhas são essenciais à produção de alimentos para o ser humano e para a manutenção dos ecossistemas.

Nos EUA, estima-se que o valor dos serviços de polinização por abelhas seja de, em média, $ 15 bilhões por ano (MORSE; CALDERONE, 2000; CALDERONE, 2012). No entanto, manter o bem estar da colônia das abelhas é um desafio crescente em muitas partes do globo. Ainda nos EUA, os apicultores relatam perdas médias de colônias no inverno de cerca de 30%, segundo a *Bee Informed Partnership* (BIP)[1], que começou a realizar pesquisas em 2010 (KULHANEK *et al.*, 2017). Muitos fatores contribuem para essas perdas, tais como: doenças, abelhas operárias fracas devido a parasitas, baixa população no inverno, abelhas operárias com reservas corporais com pouca gordura e baixos estoques de mel (POTTS *et al.*, 2016).

No Brasil, 85 das 141 espécies de plantas cultivadas para uso na alimentação humana, produção animal, biodiesel e fibras dependem em certo grau da polinização animal (GIANNINI *et al.*, 2015). Os apicultores do semiárido nordestino do país, nos últimos anos, devido às secas severas, enfrentaram graves perdas na produção de mel devido a um processo biológico conhecido como abandono da colmeia (onde a colônia abandona completamente a colmeia), característica das abelhas africanizadas criadas, em sua maior parte, na América do Sul. Nesse processo, todas as abelhas deixam o ninho estabelecido e migram para busca de um novo local mais adequado (FREITAS *et al.*, 2007). Por exemplo, na estiagem prolongada de 2012, o abandono da colmeia levou 70% das colônias a abandonarem o ninho, o que causou uma queda de 66% na produção de mel em relação a 2011 apenas no estado do Piauí (KRIDI *et al.*, 2016). Além dessa questão relativa à eficiência produtiva de mel, pode haver também diferença significativa na qualidade dos frutos gerados através de um polinização eficiente (por exemplo, a realizada pela abelhas) com relação aos que não foram polinizados por abelhas, com risco, inclusive, de má formação do fruto (vide Figura 1).

Por outro lado, no Canadá, o relatório anual de perdas do inverno de 2019, recém-lançado pela Associação Canadense de Apicultores Profissionais ou *Canadian Association of*

[1] https://bip2.beeinformed.org/loss-map/

Figura 1 – Morangos mal polinizados (à esquerda) vs bem polinizados (à direita)

Fonte: (MALAGODI-BRAGA, 2018)

Professional Apiculturists (CAPA) [2], indica um aumento no número de colônias desde 2007, mesmo com as perdas anuais de inverno. As causas das perdas no inverno, conforme relatadas pelos apicultores Canadenses, são essencialmente relacionadas ao clima e ao manejo (CURRIE *et al.*, 2010). Uma situação semelhante à do Canada pode ser observada na União Européia, conforme relatado pela Associação de Prevenção de Perdas de Colônias de Abelha ou *Prevention of honey bee COlony LOSSes Association* (COLOSS) [3] (BRODSCHNEIDER *et al.*, 2018).

1.2 Caracterização do Problema

Assim, em linhas gerais, a problemática abordada nessa tese consiste em:

Como reconhecer e predizer os níveis de bem estar das colônias de abelhas* Apis melífera? *E, em especial, como prever a perda da capacidade de termorregulação de uma colônia?

Para tentar identificar com antecedência problemas nas colônias, geralmente, o apicultor realiza um procedimento de inspeção na colônia, que consiste na verificação visual por meio da abertura da colmeia. É através da inspeção visual que o apicultor detecta uma série de problemas (SPIVAK; REUTER, 2001; VANENGELSDORP *et al.*, 2013; MUMBI *et al.*, 2014). No entanto, esse tipo de verificação consome tempo e requer habilidades e conhecimentos de

[2] http://www.capabees.com/capa-statement-on-honey-bees/
[3] https://coloss.org/core-projects/colony-losses-monitoring/

apicultura que podem levar anos para serem adquiridos (DINEVA; ATANASOVA, 2018b). Se não forem realizadas adequadamente, as inspeções podem atrapalhar o equilíbrio do microclima dentro da colmeia e o trabalho das abelhas operárias (responsáveis pelo trabalho interno do ninho). Além disso, com uma manipulação descuidada das colmeias, existe a possibilidade de matar operárias ou até de esmagar a rainha pela remoção e inserção das quadros (BENCSIK *et al.*, 2015). Se feitas corretamente, os benefícios das inspeções mais frequentes das colônias superam os riscos. Como as inspeções não devem ser realizadas durante os meses de frio ou sem alimento no campo, o monitoramento é melhor realizado por sensores durante esse período.

Além disso, os apicultores parecem não ter técnicas padronizadas de apicultura para inspeções, muitas vezes sem registros sistemáticos. Menos ainda realizam análises para determinar quais práticas são eficientes para resolver ou evitar problemas (JACOBS *et al.*, 2017). Assim, estabelecendo um padrão para avaliação do bem estar das colônias de abelha, podemos empregar o enorme potencial em usar dados de inspeção de colmeias para melhorar a apicultura em geral. Por exemplo, nos EUA, uma abordagem possível é a utilização da "Lista de verificação do bem estar da colônia", do inglês *Healthy Colony Checklist* (HCC), que ajuda a agendar tarefas, atribuir recursos e acompanhar resultados de procedimentos (LEE *et al.*, 2015; GILIOLI *et al.*, 2019).

Como alternativa, o uso e a análise de sensores nas colmeias podem ajudar a reduzir a necessidade de manipulações físicas frequentes das colmeias. Existe um interesse crescente entre os apicultores em usar sensores em colmeias, o que pode ser explicado pelo progresso da *Internet of Things* (IoT). Seja na cidade ou no campo, é possível criar a chamada apicultura de precisão (ZACEPINS *et al.*, 2015; ZACEPINS *et al.*, 2017) através da IoT e da Tecnologia da Informação como um todo (BRAGA *et al.*, 2017). Entre outras magnitudes sensoriáveis que podem ser monitoradas, é possível destacar: temperatura, umidade, concentração de dióxido de carbono (CO_2) e oxigênio nas colmeias, massa da colmeia, padrões de imagem e emissão de intensidade sonora (MEIKLE; HOLST, 2015). Também é possível quantificar características ambientais com base na localização da colmeia, como temperatura, umidade, vento e chuva (KRIDI *et al.*, 2016; MURPHY *et al.*, 2016; FLORES *et al.*, 2019). Além disso, os dados de sensores podem ser combinados aos dados de inspeções para identificação dos níveis de bem estar da colônia.

1.3 Questões de Pesquisa

Para resolver o problema de determinar o status do bem estar de uma colônia de abelhas *Apis mellifera* usando uma grande quantidade de dados gerados por sensores internos, externos e inspeções físicas, nesta tese formulamos 3 Questões de Pesquisa (**QP**) fundamentais associadas à definição e previsão do bem estar da colônia. Tais **QPs** são listadas a seguir.

QP #1: *Quais são os valores típicos de temperatura, umidade relativa e peso de uma colônia ao longo do ano e entre as estações do ano?*

QP #2: *É possível identificar níveis de bem estar de colônias de abelhas usando dados de sensores internos e externos de colônias e dados de inspeções padronizadas de apicultores?*

QP #3: *Quando um apicultor deve intervir na colmeia para evitar os problemas causados pela perda da capacidade de termorregulação em uma colônia?*

Para resolver as **QPs** apresentadas, aqui aplicamos 3 métodos de mineração de dados os quais calibram algoritmos de classificação e regressão. Assume-se que os algoritmos de classificação e regressão podem prever o nível de bem estar das colônias de abelhas com um alto grau de precisão.

1.3.1 Hipóteses

A principal hipótese de pesquisa assumida é que *um método bem definido para gerar modelos de classificação e regressão pode prever o nível de bem estar das colônias de abelhas com alto grau de precisão e baixo erro* para conjunto de dados combinados de sensores e inspeções. Para cada **QP**, foi possível associar ainda uma hipótese específica, como pode ser vista a seguir.

Hipótese #1 associada à **QP #1**: a primeira hipótese desse trabalho é que seja possível gerar modelos de classificação de dados de sensores sem dados de observações *in loco* de alta precisão e baixo erro. Esse modelos, buscam identificar faixas de valores das grandezas sensoriadas e a associação desses valores com níveis de bem estar ou fases do ciclo de vidas das colônias de abelhas.

Hipótese #2 associada à **QP #2**: a segunda hipótese assumida nessa tese é de que seja possível a utilização de inspeções *in loco* padronizadas como padrões de dados (ou classes) para a criação de modelos de classificação de dados de sensores de colônias de abelha de alta

precisão e baixo erro. As inspeções padronizadas são um mecânismo para auxiliar o apicultor a fim de gerar dados que podem ser usados para determinar o nível de bem estar, algum fenômeno ou condição específica em uma colônia.

Hipótese #3 associada à **QP #3**: a terceira hipótese assumida é de que seja possível, utilizando dados de sensores de colmeias pre-classificados em colmeias termoreguladas e colméias não termoreguladas, criar modelos de regressão de alta precisão e baixo erro para predizer a iminência da perda da capacidade de termoregulação de colônias de abelha.

1.4 Objetivo Principal e Metas

O objetivo central desta tese é *a proposição de metodologias para categorização e predição de níveis de bem estar de colônias de abelhas*. As referidas metodologias englobam as etapas de aquisição, tratamento, agrupamento, rotulagem, calibragem, classificação e predição. Para o atendimento desse objetivo, o mesmo foi decomposto nas seguintes metas.

- Meta 01: obter conjuntos de dados que compreendam ciclos anuais completos;
- Meta 02: fragmentar os conjuntos de dados em janelas de tempo;
- Meta 03: detectar e remover as anomalias dos dados;
- Meta 04: normalizar/padronizar os dados;
- Meta 05: reconhecer a quantidade ideal de classes de predição;
- Meta 06: caracterizar os níveis de bem estar de interesse;
- Meta 07: rotular os pontos amostrais de dados a partir do conhecimento de um *expert*;
- Meta 08: calibrar hiper parâmetros dos algoritmos de classificação;
- Meta 09: treinar, testar e validar algoritmos de classificação;
- Meta 10: comparar desempenho dos algoritmos de classificação;
- Meta 11: interpretar as predições realizadas;
- Meta 12: obter modelo de classificação dos níveis de bem estar de colônias.

1.5 Contribuições

Esta tese apresenta três contribuições principais. Cada uma delas busca responder uma questão de pesquisa específica da Seção 1.3 e estão apresentadas na qualidade de modelos de classificação de dados. Os modelos aplicados são indicados para dois tipos de conjuntos de dados de colmeias de abelhas compostos por dados de sensores e de inspeções, a saber:

(i) rotulado e (ii) não rotulado. Nos dados rotulados, os rótulos podem ser obtidos através de inspeções de apicultores ou através de entrevistas com apicultores. Os dados não rotulados são analizados para obtenção de padrões e posterior rotulagem. Três tipos de técnicas de mineraçã de dados específicas são usadas como base para método, a saber: (i) agrupamento, (ii) classificação e (iii) regressão.

Contribuição #1: Um método agrupamento-classificação para mineração de padrões sazonais de colônias de abelhas *Apis mellifera* (MACIEL *et al.*, 2018c; MACIEL *et al.*, 2018b; MACIEL *et al.*, 2018a) (**QP #1**).

Contribuição #2: Um método que combina dados de inspeções de colmeias, de sensores internos e de sensores atmosféricos para prever o nível de bem estar das colônias de abelhas *Apis mellifera* (BRAGA *et al.*, 2019; BRAGA *et al.*, 2019; BRAGA *et al.*, 2020) (**QP #2**).

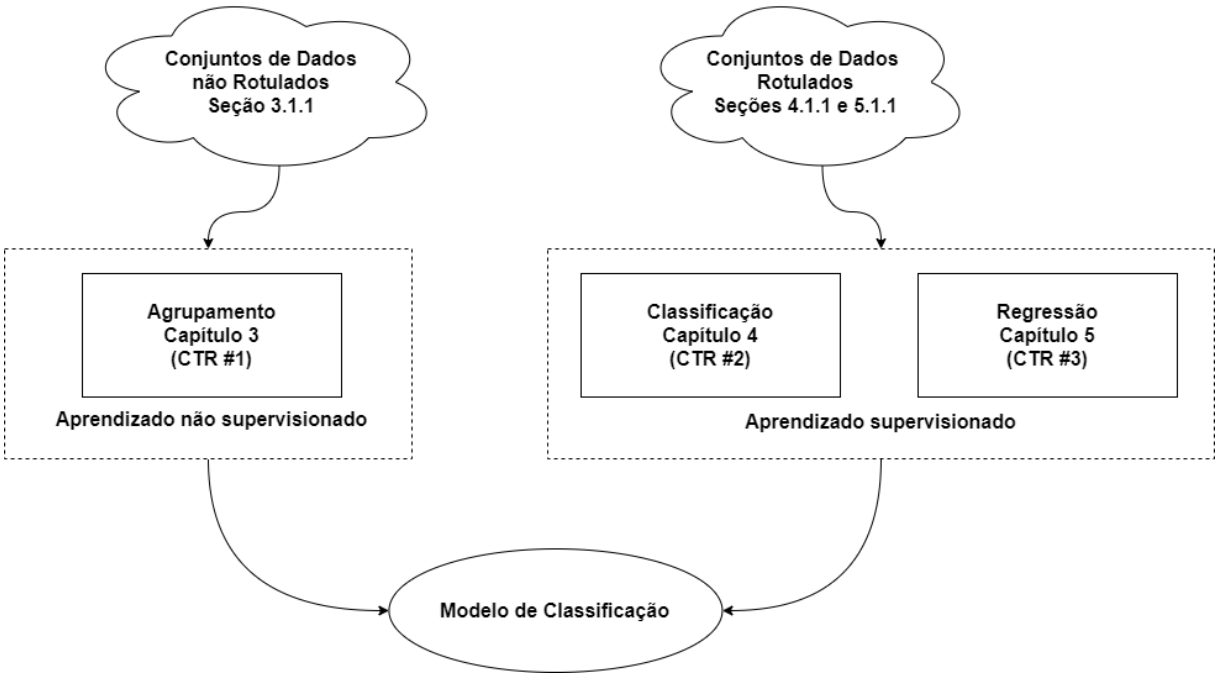
Contribuição #3: Um método que aplica o algoritmo LSTM multivariado de para prever a perda de capacidade de termorregulação em colônias de abelhas *Apis mellifera* (BRAGA *et al.*, 2019) (**QP #3**).

1.6 Organização da Tese

A Figura 2 ilustra a organização da tese com relação a utilização dos conjuntos de dados rotulado e não rotulado e as respectivas técnicas de aprendizado de máquina. Na sequência, apresenta-se a descrição dos capítulos da tese.

- **Capítulo 2**: aborda os principais temas que compõem a base teórica desta tese de doutorado. São apresentados os conceitos fundamentais relacionados a análise de dados e à apicultura. Apresenta-se ainda os principais trabalhos relacionados a esta tese;
- **Capítulo 3**: descreve a abordagem baseada em agrupamento + classificação para obtenção de um modelo de classificação de dados de colmeia. São apresentados de maneira detalhada as tarefas definidas para obtenção desse modelo, bem como os resultados obtidos, discussões e limitações dessa abordagem.
- **Capítulo 4**: descreve a abordagem baseada em classificação para obtenção de um modelo de classificação de dados de colmeia. São apresentados de maneira detalhada as tarefas definidas para obtenção desse modelo, bem como os resultados obtidos, discussões e limitações dessa abordagem.
- **Capítulo 5**: descreve a abordagem baseada em regressão para obtenção de um modelo

Figura 2 – Visão geral dos métodos aplicados.



Fonte: Elaborada pelo autor (2019)

de classificação de dados de colmeia. São apresentados de maneira detalhada as tarefas definidas para obtenção desse modelo, bem como os resultados obtidos, disc ussões e limitações dessa abordagem.

- **Capítulo 6**: apresenta as principais conclusões alcançadas bem como as limitações e possibilidades de melhorias desse trabalho.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Este capítulo aborda os principais temas que compõem a base teórica desta tese de doutorado. Na primeira seção aborda-se a apicultura de forma geral e a apicultura de precisão. Na segunda seção são apresentados os conceitos fundamentais relacionados às técnicas de análise de dados utilizadas e à apicultura. A seção seguinte apresenta os principais trabalhos relacionados, suas diferenças e semelhanças com esta tese.

2.1 Apicultura

A apicultura é a prática do manejo das abelhas *Apis mellifera* nas colmeias, principalmente para a produção de mel e para a polinização (FORMATO; SMULDERS, 2011). A prática do homem de criar abelhas é tão antiga que não se sabe ao certo quando ocorreu o princípio da apicultura (ZOGOVIC *et al.*, 2017). A história mundial da apicultura foi sendo conhecida basicamente a partir de registros pictóricos e escritos do passado, tradições orais e relatos ocasionais de testemunhas oculares de viajantes que visitavam diferentes partes do mundo (CRANE, 1999). Nas primeiras abordagens do cultivo de abelhas, o homem chegava a destruir as colônias para poder ter acesso aos produtos gerados pelos insetos. Com o passar do tempo e a aquisição de conhecimento, as técnicas de manejo melhoraram, e a apicultura passou a ter outros objetivos, como tirar proveito da função de polinização desempenhada pelas abelhas (ZOGOVIC *et al.*, 2017).

Nos últimos 200 anos, a prática do cultivo de abelhas sofreu um desenvolvimento mais intenso, com o reconhecimento de propriedades importantes das colônias por parte dos apicultores e a modernização das colmeias (ZOGOVIC *et al.*, 2017). De longe, a contribuição mais importante que as abelhas fazem para a agricultura moderna são os serviços de polinização que eles fornecem (VANENGELSDORP; MEIXNER, 2010). Com isso, a apicultura se tornou mais eficiente, e a prática contemporânea passou a ser descrita como apicultura racional, na qual o apicultor retira os produtos das colmeias incomodando minimamente as abelhas (ZOGOVIC *et al.*, 2017).

2.1.1 Apicultura de Precisão

Ainda no século 20 surgiu o interesse de se monitorar e coletar dados de colônias de abelhas. Há registro de coleta de dados de temperatura e massa de uma colmeia por diversos dias

entre 1907 e 1908 (MEIKLE; HOLST, 2015). Atualmente, com a modernização da comunicação e dos sensores, e com a tecnologia da informação cada vez mais presente em todas as áreas do conhecimento, apicultores e pesquisadores podem monitorar remotamente vários aspectos físicos das colônias de abelhas (MEIKLE; HOLST, 2015; ZOGOVIC *et al.*, 2017). Essa é a premissa da apicultura de precisão, um sub-ramo da Agricultura de Precisão (ZACEPINS *et al.*, 2015).

A apicultura de precisão envolve basicamente as seguintes etapas: coletar individualmente os dados das colmeias, analisar a informação obtida e dar suporte à tomada de decisão para o gerenciamento das atividades em um apiário. Zacepins (2012) definiu a apicultura de precisão como uma estratégia de gerenciamento de apiários baseada no monitoramento de colônias de abelhas para minimizar o consumo de recursos e maximizar a produtividade das abelhas. Essa abordagem visa reduzir o desperdício de recursos e o estresse das abelhas causado por atividades desnecessárias (ZACEPINS *et al.*, 2015). Uma vez que os sensores são instalados, as colônias podem ser monitoradas sem perturbação, inclusive durante os períodos quando as inspeções invasivas são contraindicadas, como durante o período frio (MEIKLE; HOLST, 2015).

De acordo com Zacepins *et al.* (2015), a etapa de coleta de dados pode ser classificada em três níves de informações possíveis, a saber:

1. **parâmetros de nível de apíário**: incluem parâmetros meteorológicos, de oferta de alimentação para a colônia e de localização da colônia ou até mesmo através de camêras de vídeo monitoramento, dentre outros; pode explicar algumas peculiaridades do comportamento das colônias. Por exemplo, o aumento do número de abelhas entrando e saindo pode ser causado por um ruído externo, que podem ser registrados pelas câmeras de vídeo usadas para observar o apiário. Um aumento no peso da colméia pode ser causado por chuva que pode ser detectada por vídeo ou por uma estação meteorológica (ZACEPINS *et al.*, 2015).
2. **parâmetros de nível de colônia**: temperatura, umidade, concentração de gases, sons, vídeos, vibrações e peso da colmeia, dentre outros; atualmente, esse é o métodos mais popular de coleta de dados relacionados às colmeias. Cada parâmetro físico de monitorado dentro da colônia ou a composição deles pode explicar um determinado fênomeno ou conjunto de fenômenos da colônia. Assim, os fenômenos que ocorrem dentro da colmeia possuem uma grande complexidade e a quandidade de sensores instalados ajuda sobremaneira à detecção e previsão desse fenômenos. A exemplo da temperatura, umas das grandezas mais importantes, está diretamente associada com a termorregulação. O peso com a produção de mel e a luminosidade interna com a saída das abelhas para o pasto

apícola.

3. **parâmetro relacionados às abelhas individualmente**: número de abelhas entrando e saíde da colmeia, número de abelhas a entrada da colmeia, número aproximado de abelhas dentro da colmeia compondo as varias 'castas' do enxame, dentre outros; objetiva observar se as abelhas estão individualmente ao redor da entrada da colméia. Pode ser implementado através de duas estratégias gerais: (i) monitoramento por vídeo e (ii) contadores de abelhas. Essa informações são importantes que verificar se a colônia está composta adequatamente e se as abelhas estão realizando as tarefas habituais da colônia corretamente, tais como: saída para o pasto apícola e proteção da colônia.

2.2 Descoberta de Conhecimento em Banco de Dados

A metodologia ou processo para descoberta de conhecimento em banco de dados mais amplamente utilizada é o processo chamado *Descoberta de Conhecimento em Banco de Dados*, em inglês, *Knowledge Discovery in Databases* (KDD) (PIATETSKY-SHAPIRO, 1990; FAYYAD *et al.*, 1996). De acordo com Fayyad *et al.* (1996), o processo KDD pode ser definido, em um nível mais abstrato, como o campo que se preocupa com o desenvolvimento de métodos e técnicas para entender os dados. O KDD é um processo interativo e iterativo, que envolve etapas com decisões tomadas pelo usuário.

Uma ilustração dos passos propostos pelo KDD pode ser vista na Figura 3. Como se vê, a partir de dados brutos é possíve extrair conhecimento desse dados realizando-se cinco atividades, a saber: (i) **seleção** - seleção um conjunto de dados ou um subconjunto de variáveis ou pontos amostrais; (ii) **preprocessamento** - remoção de ruído e definição estratégias para lidar com dados ausentes; (iii) **transformação** - redução ou transformação de dimensionalidade e de seleção do número efetivo de variáveis; (iv) **mineração de dados** - faz uso de método de mineração de dados, por exemplo: classificação, regressão ou agrupamento; e (v) **interpretação** - interpretação dos padrões minerados, possivelmente retornando a qualquer uma das etapas anteriores para iterações adicionais. Pode envolver também a visualização dos padrões e modelos extraídos ou a visualização dos dados.

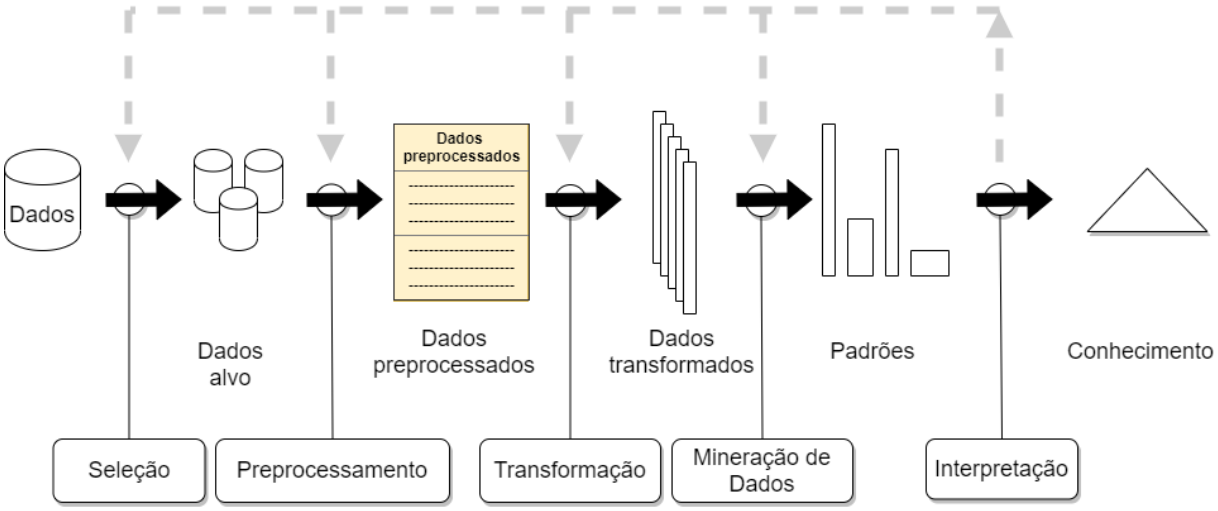
A mineração de dados, em particular, se refere ao processo de extração de informação implícita, previamente desconhecida e potencialmente útil de um conjunto de dados (PIATESKI; FRAWLEY, 1991). Ela pode ser baseada em diversas abordagens. Há, por exemplo, a mineração baseada em técnicas de reconhecimento de padrões, a mineração baseada em teorias estatísticas

e matemáticas, e as abordagens integradas (CHEN *et al.*, 1996). O propósito da abordagem baseada em reconhecimento de padrões é realizar o aprendizado de um padrão ou de maneira supervisionada ou de maneira não-supervisionada.

No aprendizado supervisionado o padrão de entrada é identificado como membro de uma classe predefinida, enquanto que no aprendizado não-supervisionada o padrão é atribuído a uma classe desconhecida (JAIN *et al.*, 2000). Neste caso, como não há informação prévia sobre as classes às quais os dados pertencem, utiliza-se apenas as informações extraídas dos próprios dados para tentar agrupá-los por similaridade. Na aprendizagem não-supervisionada, o agrupamento é considerada a abordagem mais importante (XU; TIAN, 2015). Na aprendizagem supervisionada, a classificação e a regressão são as abordagens mais utilizadas.

Esta tese apresenta três médotos de análise de dados que são ***instâncias*** da metodologia KDD, ou seja, foram construídas tomando como base o processo apresentado na Figura 3. Como apresentado no Capítulo 1, os métodos são categorizados de acordo com a abordagem de reconhecimento de padrões utilizada, ou seja, aprendizado supervisionado ou não-supervisionado. Para o aprendizado supervisionado, apresenta-se dois métodos, um que realiza a mineração de dados com a classificação e outro com a regressão. Para o aprendizado não-supervisionado, apresenta-se um método que realiza a mineração de dados com o aprupamento.

Figura 3 – Método de predição via aprendizado não supervisionado



Fonte: Adaptado de Fayyad *et al.* (1996)

2.2.1 Agrupamento

O Agrupamento, ou clusterização, tem como objetivo separar um conjunto finito de dados não rotulados em um conjunto finito e discreto de estruturas de dados "naturais" ocultas (XU; WUNSCH, 2005). A seguir, tem-se a definição clássica do agrupamento dada por Jain e Dubes (1988):

- instâncias no mesmo grupo, ou *cluster*, devem ser tão semelhantes quanto possível;
- instâncias em grupos diferentes devem ser tão diferentes quanto possível;
- a medida de similaridade e dissimilaridade deve ser clara e ter significado prático.

Entre os algoritmos de agrupamento mais tradicionais estão os algoritmos baseados em partições, que têm como ideia básica considerar o centro dos dados como o centro do grupo correspondente (XU; TIAN, 2015).

O algoritmo de agrupamento baseado em partições mais utilizado na prática é o *k-means* (WU *et al.*, 2008), cujo conceito é recalcular o centro de cada grupo por um processo iterativo que irá continuar até que um critério de convergência seja atendido (XU; TIAN, 2015). Este algoritmo é eficaz em conjuntos de dados de larga escala (WU *et al.*, 2008), e sua complexidade aumenta de forma aproximadamente linear com o número de pontos amostrais do conjunto de dados (XU; WUNSCH, 2005).

2.2.1.1 Validação do Agrupamento

Em um problema de agrupamento, é comum não conhecer *a priori* o valor k que melhor se ajusta ao conjunto de dados, e o algoritmo *k-means* requer esse valor antes que ele possa ser executado. Uma maneira de encontrar o valor ideal de k é através dos índices de validação de grupos (CALINSKI; HARABASZ, 1974). Normalmente, esses índices avaliam dois aspectos de um grupo: (*i*) a coesão interna, com base na distância entre os vetores de dados do mesmo grupo, o que indica o quão compacto é um grupo; (*ii*) e a separação externa, com base nas distâncias entre os centróides dos grupos que indicam a distinção dos grupos.

Milligan e Cooper (1985) avaliaram e compararam 30 índices de acordo com seus desempenhos em relação a uma série de conjuntos de dados artificiais, e o índice *Calinski-Harabasz* (CH) (CALINSKI; HARABASZ, 1974) obteve os melhores resultados. Para determinar o valor de k, neste trabalho, usamos o índice CH (CALINSKI; HARABASZ, 1974). O cálculo desse índice é definido por (2.1), onde B_k e W_k são, respectivamente, as matrizes de dispersão

entre grupos e intragrupo, a partir dos quais os traços são calculados (operador $tr()$). N é o número de pontos amostrais para cada grupo.

$$CH(K) = \frac{tr(B_k)/(k-1)}{tr(W_k)/(N-k)}. \tag{2.1}$$

O índice *CH* deve ser calculado para vários valores de k e, como um valor alto de B_K (grupos muito distintos) e um valor baixo de W_K (grupos compactos) são desejados, o valor mais alto de *CH* indica o k mais apropriado para ser o número de grupos.

2.2.1.2 k-means

O algoritmo de agrupamento usado neste trabalho foi o *k-means* (MACQUEEN, 1967) e o número ideal k é obtido através de uma análise exploratória com validação de agrupamento. O objetivo do *k-means* é dividir o conjunto de vetores de dados N em partições (ou grupos, ou *clusters*) k sem sobreposição ($k \ll N$), com o auxílio de protótipos de k, também chamados de centróides, posicionados corretamente no espaço de dados. Então, cada vetor de dados é associado a um centróide por critério de similaridade, por exemplo, a menor distância.

O conjunto W de k centróides é representado por (2.2), onde p é o tamanho do vetor de dados.

$$W = \{w_i\}_{i=1}^{k} \mid w_i \in \mathbb{R}^p. \tag{2.2}$$

O grupo V associado a cada centróide w é definido por (2.3), onde x é um vetor de atributo e $\|x-w\|$ indica a distância euclidiana.

$$V_i = \{x \in \mathbb{R}^p \mid \|x-w_i\| < \|x-w_j\|, \forall j \neq i\}. \tag{2.3}$$

Em geral, o *k-means* se comporta da seguinte maneira: inicialmente seleciona-se vetores aleatórios k no espaço de dados como protótipos; o grupo de cada protótipo w_i é determinado por 2.3; o algoritmo calcula a nova posição de cada protótipo w_i como a média dos objetos de um grupo N_i V_i (Eq.2.4).

$$w_i = \frac{1}{N_i} \sum_{x \in V_i} x. \tag{2.4}$$

O grupo V_i (2.3) e a nova posição do protótipo w_i (2.4) são recalculados repetidamente até a convergência do algoritmo. Isso acontece quando a posição do protótipo w_i (2.4) não muda mais ou quando atinge um número máximo de iterações, por exemplo. Para avaliar quantitativamente o posicionamento dos protótipos, calculamos a Soma da Distância Quadrada, do inglês, *Sum of Squared Differences* (SSD) de um vetor de dados até o centróide mais próximo (2.5). Este erro ajuda a avaliar a qualidade dos grupos gerados.

$$SSD = \sum_{\forall x \in V_i} \|x - w_i\|^2. \tag{2.5}$$

As principais limitações do *k-means* são sua sensibilidade à inicialização, já que os centros iniciais são atribuídos aleatoriamente, e às anomalias do conjunto de dados, já que os centróides obtidos pelo algoritmo são gerados a partir da média dos pontos amostrais do conjuntos de dados, e esta não é uma medida estatística robusta (WU *et al.*, 2008). Isso torna indispensável a adoção da estratégia de remoção de anomalias como uma etapa de pré-processamento dos dados. Além disso, o *k-means* não é eficaz ao manipular conjuntos de dados com muitos atributos (XU; WUNSCH, 2005).

2.2.2 Classificação

Neste trabalho, a classificação foi utilizada para criar um método capaz de identificar estados de novos pontos amostrais coletados pelo sistema de monitoramento. A partir das classes definidas na fase de agrupamento e validadas pelo especialista ou a partir de inspeções padronizadas, um método de classificação com alta precisão pode prever em tempo real o estado das colmeias, permitindo a emissão de alertas.

A seguir, os algoritmos de classificação usados neste trabalho são descritos. Cinco abordagens foram escolhidas: *Naive Bayes* (NB) (MARON, 1961), um classificador probabilístico que funciona com base no teorema de Bayes; o k-Vizinhos mais Próximos ou *k-Nearest Neighbor* (kNN) (COVER; HART, 1967), que define os exemplos de treinamento mais próximos k; as Florestas Aleatórias ou *Random Forest* (RF) (HO, 1995), que usa uma coleção de árvores de decisão e as Redes Neurais ou *Neural Networks* (NN) (MCCULLOCH; PITTS, 1943), que é baseado em uma metáfora do comportamento do cérebro; e as Máquinas de Vetores de Suporte ou *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995).

2.2.2.1 *Naive Bayes*

O classificador NB é baseado no Teorema de Bayes (2.6) para gerar as previsões para cada observação, classificando um ponto amostral em um grupo com maior probabilidade de ter seus atributos.

$$P(Y \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid Y) \times P(Y)}{P(\mathbf{X})}. \tag{2.6}$$

Formalizando o problema de classificação em termos estatísticos, a variável X indica o conjunto de pontos amostrais e Y a variável de classe. Assumindo uma relação não determinística entre essas variáveis, a probabilidade condicional $P(Y|X)$ é conhecida como probabilidade posterior de Y e $P(Y)$ como sua probabilidade anterior.

A probabilidade anterior $P(Y)$ pode ser avaliada a partir do conjunto de treinamento, calculando a fração de registros pertencentes a cada classe. Supondo que os atributos sejam condicionalmente independentes, um classificador Bayes simples avalia a probabilidade condicional da classe $P(X_i)$, dada a etiqueta de classe y (2.7), na qual cada conjunto $\mathbf{X} = \{X_1, X_2, ..., X_d\}$ consiste em d atributos.

$$P(\mathbf{X} \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y). \tag{2.7}$$

Assim, em vez de calcular a probabilidade condicional da classe para cada combinação de X, basta estimar apenas a probabilidade condicional de cada X_i, dado Y. Portanto, para classificar um novo registro, basta calcular a probabilidade posterior para cada classe Y (2.8).

$$P(Y \mid \mathbf{X}) = \frac{P(Y) \prod_{i=1}^{d} P(X_i \mid Y)}{P(\mathbf{X})}. \tag{2.8}$$

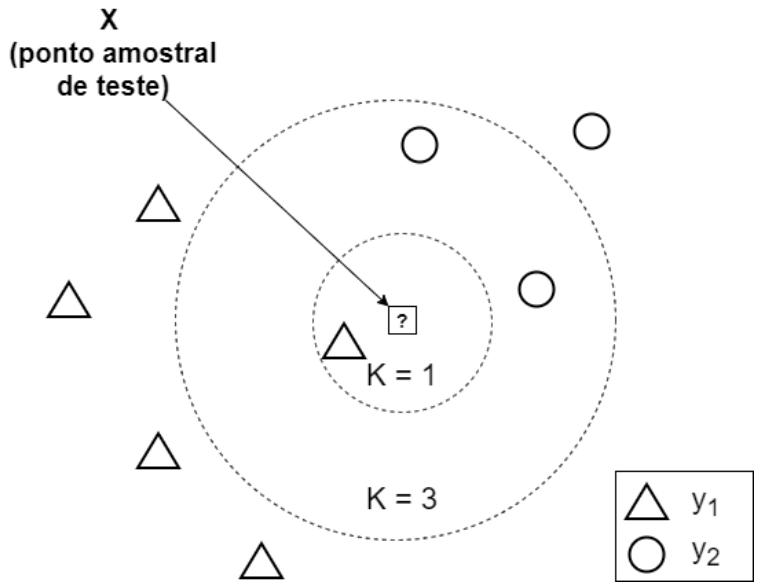
Considerando que $P(\mathbf{X})$ é fixo para cada Y, basta escolher a classe que maximiza o termo numerador, $P(Y) \prod_{i=1}^{d} P(X_i \mid Y)$.

2.2.2.2 *k-Vizinhos mais Próximos*

O classificador k-Vizinhos mais Próximos, ou em inglês kNN é baseado no recebimento de um conjunto de treinamento, aprendendo com esse conjunto, validando o aprendizado com um conjunto de testes e, ao receber novas observações, classificando-os de acordo com os

conhecidos (TAN *et al.*, 2005). Cada nova observação tem sua distância calculada com cada observação já conhecida. A classificação é realizada de acordo com o maior número de vizinhos mais próximos de k pertencentes à mesma classe. Uma aproximação inicial ao valor de k pode ser dada pela raiz quadrada do número de observações presentes no conjunto de dados. A Figura 4 apresenta uma ilustração do processo de classificação com o kNN.

Figura 4 – Modelo básico do k-Vizinhos mais Próximos



Fonte: Elaborada pelo autor (2020)

No exemplo da Figura 4, o ponto amostral de teste (quadrado) deve ser classificado em triangulo (y_1) ou em circulo (y_2). Se $k = 1$ (círculo interno), o quadrado, então, é atribuído à classe triângulo (y_1) porque existe 1 triângulo dentro do círculo interno. Se $k = 3$ (círculo externo), ele é atribuído ao circulo (y_2), pois, nesse caso, tem-se 2 círculo vs. 1 triângulo dentro do círculo externo.

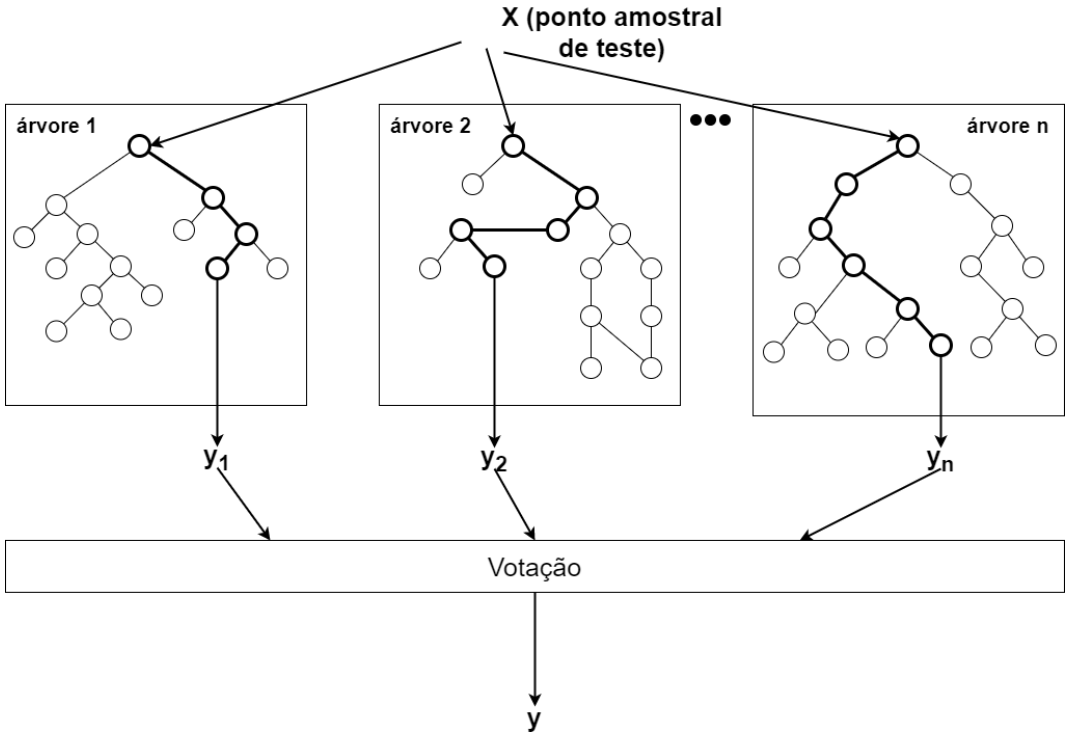
A escolha final do valor de k, por sua vez, não é tão simples. Se um k muito pequeno for escolhido, as previsões podem ser distorcidas por alguns vizinhos e podem gerar classificações incorretas. Por outro lado, se k for muito grande, a classificação poderá acabar invadindo regiões de outras classes e também gerar classificações incorretas. Para superar essa dificuldade, a técnica *validação cruzada* será usada para selecionar um valor justo de k.

2.2.2.3 Florestas Aleatórias

O classificador RF baseiam-se em árvores de decisão para gerar suas classificações. Árvores de decisão são estruturas que baseiam-se em regras de decisão para se ramificarem em possibilidades e criar um "caminho". No final do caminho está a classificação atribuída à entrada.

As árvores de decisão têm como desvantagem a sua instabilidade, pois uma pequena alteração no conjunto de treino pode gerar uma árvore totalmente diferente e, consequentemente, predições equivocadas. Para aumentar a acurácia e diminuir a instabilidade das árvores de decisão, o algoritmo do RF utiliza várias árvores em conjunto, criando uma espécie de floresta. Uma floresta é mais estável e menos suscetível a pequenas alterações. A Figura 5 apresenta uma ilustração do processo de classificação com o algoritmo de árvores de decisão.

Figura 5 – Modelo básico de uma árvore de decisão



Fonte: Elaborada pelo autor (2020)

No processo de aprendizagem do algoritmo, são criadas as regras que definem os ramos da árvore, e a partir daí desenvolve-se uma floresta com várias árvores. O processo de criação das árvores acontece de forma que elas possuam baixa correlação, a fim de reduzir o custo computacional e aumentar a eficácia do classificador. Para tal, cada árvore é montada

usando um número *m* de preditores escolhidos aleatoriamente entre os *n* preditores originais (*m* é o mesmo para todas as árvores). No processo de classificação de novos pontos amostrais, cada árvore irá realizar a classificação individualmente do ponto amostral **X**. Então, através de um processo de votação, escolhe-se para **X** a classe mais votada entre as árvores.

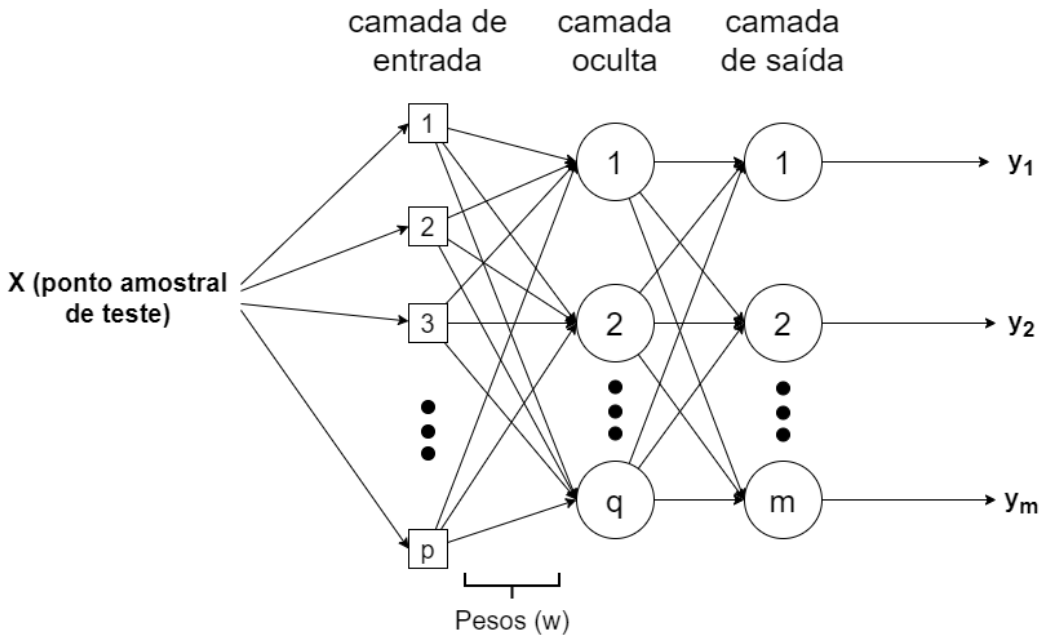
2.2.2.4 Redes Neurais

O classificador NN baseiam-se em uma metáfora do comportamento do cérebro. São formadas por unidades de processamento simples denominados neurônios que são responsáveis pelo cálculo de determinadas funções matemáticas. Essa abordagem foi proposta inicilmente por McCulloch e Pitts (1943). Os neurônios ficam organizados em uma ou mais camadas e interligados por conexões (sinapses) com pesos. Das diversas topologias existentes para os neurônios, uma bastante conhecida é a rede Perceptron Simples (PS) (ROSENBLATT, 1958), onde multiplas entradas são totalmente conectadas a uma única camada, cujos neurônios determinam a saída da rede.

A rede Perceptron Multicamadas, do inglês, *Multi Layer Perceptron* (MLP), utilizada nesse trabalho, é obtida a partir da rede PS e pode ser usada em problemas de classificação não-lineares. Na MLP, é possível destacar (i) as unidades de entrada, responsáveis pela simples passagem dos valores de entrada para os neurônios das camadas seguintes, (ii) a(s) camada(s) oculta(s), que contêm neurônios responsáveis pelo processamento não-linear da informação de entrada e (iii) a camada de saída, que contêm neurônios responsáveis pela geração da saída da rede neural. A Figura 6 ilustra as camadas supracitadas.

Na utilização das MLP's em tarefas de classificação de padrões, deve-se associar a cada padrão de entrada (ponto amostral **X**, Figura 6) uma das classes predefinidas. Assim, uma MLP com uma camada oculta pode representada por: $MLP(p,q,m)$, onde, p é o número de variáveis de entrada, q é o número de neurônios ocultos, m é o número de neurônios de saída. A depender a quantidade de camadas ocultas que uma rede MLP possui, defini-se a arquitetura da rede. Por exemplo, a arquitetura 8:2:5:6 possui 8 neurônios na camada de entrada, duas camadas ocultas (com 2 e 5 neurônios) e uma camada de saída com 6 neurônios. O processamento de cada neurônio se dá através da função de ativação, que é responsável por analisar o sinal gerado pela combinação linear das unidades de entrada e dos pesos das sinapses (w), para gerar o sinal de saída do neurônio.

Figura 6 – Modelo básico de uma rede perceptron multicamadas



Fonte: Elaborada pelo autor(2020)

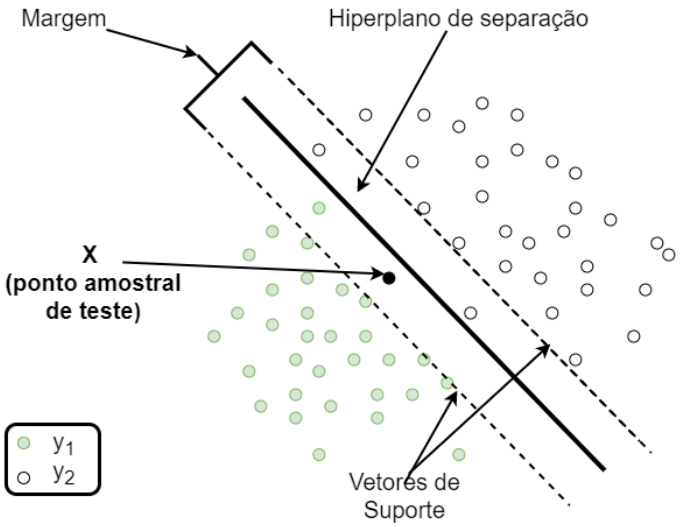
2.2.2.5 Máquinas de Vetores de Suporte

O classificador SVM é uma extensão dos classificadores de vetores de suporte, onde o espaço real dos dados é "aumentado" pela utilização dos *kernels*. O "Truque do Kernel" ou o chamado "*kernel trick*" é a generalização de um produto interno de dois vetores $\langle a, b \rangle$ na obtenção da equação final do algoritmo. Com um kernel linear o produto é o mesmo, porém, com os kerneis *Radial Basis Feature* (RBF) e *Polynomial*, por exemplo, sobre esse produto interno é aplicado uma função g, levando a um espaço aumentado e, consequentemente, adicionando de caracteristicas não-lineares ao hiperplano de decisão. Para predizer a qual classe uma determinada observação x_0 pertence, o classificador efetua o produto interno aplicado à função g e atribui a classe tomando como base o hiperplano de decisão.

2.2.3 Regressão

Neste trabalho, a regressão foi utilizada para criar um método capaz de identificar a perda da capacitadade de termoregulação de uma colônia de abelhas. A partir de uma pré-classificação realizada por um especialista, um modelo de classificação com alta precisão foi obtido.

Figura 7 – Modelo básico de uma máquina de vetores de suporte



Fonte: Elaborada pelo autor (2020)

2.2.3.1 Long Short-Term Memory (LSTM)

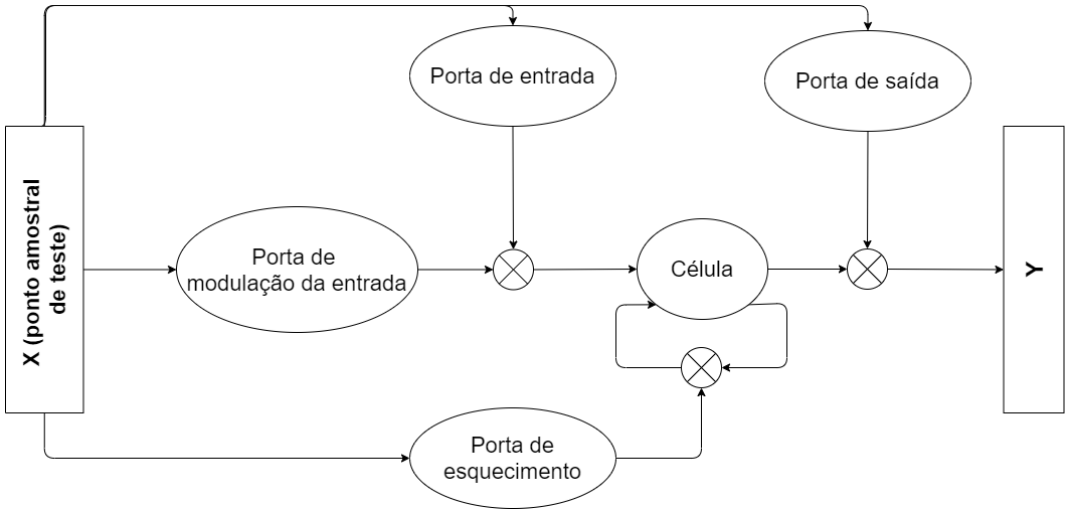
Uma Rede Neural Recorrente, do inglês *Recurrent Neural Network* (RNN), é um tipo de rede neural artificial projetada para reconhecer padrões em seqüências de dados. Para isso, eles levam em conta os dados percebidos anteriormente no tempo. Eles fazem isso através do loop de *feedback* conectado às suas decisões passadas, processando as próprias saídas momento após momento como entrada. Costuma-se dizer que as redes recorrentes têm memória.

O LSTM é uma arquitetura RNN específica que foi projetada para modelar seqüências temporais e suas dependências de longo alcance com mais precisão do que as RNN convencionais (SAK *et al.*, 2014). Como qualquer Rede Neural, ela possui neurônios e conexões, cada conexão tem um peso calculado através da retroalimentação ou *Back Propagation*. O *back propagation* envia o erro final para as saídas, pesos e entradas de cada camada oculta, atribuindo a esses pesos a responsabilidade por uma parte do erro e recalculando pesos diferentes até encontrar a composição que minimiza o erro.

Em uma rede neural, a camada mais à esquerda da rede é chamada de camada de entrada e a camada mais à direita, a camada de saída. A camada intermediária dos nós é chamada de camada oculta, porque seus valores não são observados no conjunto de treinamento. Na arquitetura LSTM, a camada oculta possui células LSTM, ou **LSTM cells**. Como mostra a Figura 8, cada célula possui 3 portas: uma porta de entrada (controla se a célula de memória está sendo atualizada), uma porta de esquecimento (controla se a célula é redefinida para zero) e uma porta de saída (controla se as informações da célula estão visíveis). Dessa forma, as

informações podem ser armazenadas, gravadas ou lidas em uma célula, como os dados na memória do computador.

Figura 8 – Célula de memória do LSTM



Fonte: Elaborada pelo autor (2019)

Em uma arquitetura LSTM, ainda é possível usar as camadas de *dropout* e densa. A camada de *dropout* é responsável pela execução da regularização para evitar o excesso de ajustes, evitando co-adaptações nos dados de treinamento (SRIVASTAVA *et al.*, 2014). A camada densa que é uma camada totalmente conectada, significando que todos os neurônios nessa camada estão conectados aos da próxima camada.

2.3 Revisão da Literatura

O interesse pelas abelhas e seus hábitos aumentou consideravelmente ao longo dos anos (WINSTON, 1980; MCNALLY; SCHNEIDER, 1992; STALIDZANS *et al.*, 2002; ZACEPINS, 2012; CAVALCANTE *et al.*, 2018). Algumas pesquisas se concentraram nos recursos dos sistemas de monitoramento (ZACEPINS; KARASHA, 2012; ZACEPINS *et al.*, 2016; ZACEPINS *et al.*, 2017), enquanto outras também exploram a análise dos dados monitorados nas colmeias pelos sistemas (KRIDI *et al.*, 2016; KVIESIS; ZACEPINS, 2016). No entanto, o reconhecimento de níveis de bem estar de uma colônia é de fundamental importância para o apicultor, pois com essas informações ele é capaz de usar suas colônias de forma mais produtiva em serviços de polinização e na produção de produtos apícolas. Além disso, poder executar o gerenciamento das colmeias com mais eficiência. Assim, a definição de procedimentos padro-

nizados é vital para o reconhecimento confiável do nível de bem estar de uma colônia (EFSA Panel on Animal Health and Welfare (AHAW), 2016; GILIOLI *et al.*, 2019) e para a análise subsequente dos dados coletados. A seguir, serão apresentados alguns trabalhos relacionados que realizaram a identificação do nível de bem estar ou de outros estados das colônias. Esses documentos foram analisados observando 5 características principais (i) o uso de técnicas de aprendizado de máquina, (ii) o uso de anotações de inspeção, (iii) o uso de sensores em colmeias, (iv) o uso de sensores em colmeias e sensores externos, e (v) o uso de sensores internos e externos e anotações de inspeções.

É possível encontrar muitos trabalhos na literatura que apresentaram propostas para identificar alguns fenômenos específicos que ocorrem no ciclo de vida de uma colônia. No entanto, poucos trabalhos são dedicados à identificação do estado geral do bem estar de uma colônia, embora o estado geral do bem estar seja o que o apicultor procura saber imediatamente em uma inspeção, através da identificação de pragas, número de abelhas, entre outras características. Em Lee *et al.* (2015) os autores propuseram um mecanismo baseado em questionários para identificação de doenças. Questionários ou inspeções são comuns, mas são invasivos (porque implicam a abertura da colmeia), não são precisos (dependem da subjetividade do apicultor) e não são proativos (pois geralmente só veem o problema, mas fazem pouco para ajudar a previsão). Mesmo assim, as inspeções são fundamentais para que o apicultor possa conhecer suas colônias porque é um mecanismo mais simples, barato e difundido.

Assim, inspeções padronizadas podem ser um procedimento fundamental no reconhecimento do nível de bem estar de uma colônia. Em Jacobs *et al.* (2017), os autores realizaram um estudo que perguntou a 697 apicultores nos EUA quais itens eles consideravam mais importantes em uma inspeção. O objetivo do estudo foi descobrir as melhores práticas de gestão para todos os apicultores. A pesquisa apresentada para os apicultores foi organizada de acordo com o HCC proposto por Rogers (2017). O HCC de Roger baseia-se em 6 itens principais de inspeção, a saber: (i) presença de todas as fases de cria, (ii) abelhas adultas suficientes, (iii) presença de uma rainha jovem, (iv) alimento em quantidade suficiente, (v) ausência estressor aparente e (vi) espaço adequado. Os autores observaram que pelo menos 80 % dos apicultores mencionam a importância de cria e dos ovos, rainha, armazenamento de alimentos e doenças/estressores. Além dos já mencionados, abelhas adultas e o espaço tiveram 67% e 50% de importância, respectivamente. Uma outra lista de verificação que pode ser citada são as 5 perguntas de Hooper, do inglês, *Hooper's 5 Questions* (H5Q), muito populares no Reino Unido. As cinco perguntas do

H5Q são: (i) A colônia tem espaço suficiente? (ii) A rainha está presente e deposita a quantidade esperada de ovos? (iii) a. (no início da estação) A colônia está crescendo em tamanho tão rápido quanto outras colônias no apiário? b. (meio da estação) Há células de rainha presentes na colônia? (iv) Existem sinais de doença ou anormalidade? (v) A colônia tem reservas suficientes para durar até a próxima inspeção? (HOOPER, 1996). Portanto, as duas listas de verificação observam os mesmos itens, a saber: espaço, rainha, cria, estressores e comida. Contudo, no HCC Rogers (2017) criou um novo item relacionado à cria para monitorar as abelhas adultas.

O uso de sensores também pode ser uma peça fundamental na previsão de estados indesejados em uma colônia. Os sensores, juntamente com a IoT, estão se tornando cada vez mais comuns na apicultura, criando a chamada apicultura de precisão (FLORES *et al.*, 2019). Temperatura e umidade têm sido as métricas mais usadas na apicultura de precisão (ZACEPINS *et al.*, 2015; MEIKLE; HOLST, 2015; MEIKLE *et al.*, 2017). Além de temperatura, umidade, peso, vibrações, áudio, CO_2 (dióxido de carbono), O_2 (oxigênio), gases poluentes, dados climáticos (temperatura externa, precipitação e intensidade da luz solar) podem ser usados para classificar os estados das colônias (MURPHY *et al.*, 2016). A variável umidade, por exemplo, está fortemente associada ao resfriamento por evaporação, um recurso fundamental para as abelhas controlarem a hipertermia da colmeia (OSTWALD *et al.*, 2016). O peso da colmeia também é uma métrica útil para monitorar a produtividade de uma colônia e a correlação entre a produção de mel e diferentes parâmetros para condições meteorológicas (FITZGERALD *et al.*, 2015; RUAN *et al.*, 2017). De modo mais abrangente, temperatura, umidade relativa e peso podem ser combinados para verificar a termorregulação e a evolução da colônia na estação da floração (produção de mel) (GIL-LEBRERO *et al.*, 2017). Em Seritan *et al.* (2018), é possível observar o uso dessas tecnologias para detectar o nível de bem estar de uma colônia. Os autores usaram sensores de temperatura (internos e externos à colônia), umidade, dióxido de carbono (CO_2) e peso. Os autores fizeram uma correlação entre os valores lidos pelos sensores e os valores referenciais relatados indicados na literatura, mas eles não apresentam um procedimento automático para identificar o bem estar das abelhas.

As abelhas rainhas produzem um ruído agudo específico antes de deixar a colmeia, enquanto o ruído produzido pelo resto da colônia se torna mais alto. Portanto, os sensores de áudio ajudam a detectar o estado iminente de abandono de uma colmeia (MURPHY *et al.*, 2015; MURPHY *et al.*, 2015). As imagens também podem ser uma métrica útil na tomada de decisões em um apiário. Por exemplo, um sistema de monitoramento com câmeras posicionadas na

entrada da colmeia pode capturar informações sobre o fluxo de abelhas e, assim, registrar o nível de atividade das abelhas ao redor da colmeia e ser um bom indicador do bem estar da colônia (TASHAKKORI; GHADIRI, 2015). Em relação às infestações de combate, algoritmos de processamento de imagens podem ser aplicados a vídeos internos de uma colônia, por exemplo, para detectar e até eliminar com lasers o ácaro varroa (CHAZETTE *et al.*, 2016).

Também vale a pena mencionar o uso de algoritmos de aprendizado de máquina, capazes de "aprender" através de conhecimentos anteriores e fazer previsões. No entanto, o uso desses algoritmos implica uma grande quantidade de dados (WALTON *et al.*, 2016). De tal maneira que o algoritmo é capaz de aprender como as mais diversas combinações possíveis dos dados de entrada podem implicar em uma saída que se deseja obter. Algumas técnicas de análise de dados estão sendo usadas para estudar eventos específicos, por exemplo, o agrupamento, tem sido usada com o objetivo de identificar alguns fenômenos de colmeias, como termorregulação (KRIDI *et al.*, 2016) e de padrões sazonais (STALIDZANS; BERZONIS, 2013; MACIEL *et al.*, 2018a). A Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA) e a Análise de Função Discriminante, do ingês *Discriminant Function Analysis* (DFA) foram aplicadas para identificar, respectivamente, um enxameamento (BENCSIK *et al.*, 2011) e o início da criação de uma nova cria (BENCSIK *et al.*, 2015) através de vibrações da colmeia. A Análise de Variância, do ingês *Analysis of Variance* (ANOVA) (FISHER, 1918) foi utilizada para analisar a correlação entre temperatura e a umidade com o aparecimento de doenças na colônia ao longo das estações do ano e sua exposição a áreas de pasto apícola naturais ou comerciais exploradas para polinização (MEIKLE *et al.*, 2017). Além disso, o aprendizado supervisionado tem sido usado para detectar anomalias de comportamento (CARVALHO *et al.*, 2018), produção de crias e enxameamento (KVIESIS; ZACEPINS, 2016) através de redes neurais, presença da rainha (ROBLES-GUERRERO *et al.*, 2019) através de Regressão Logística do Tipo Lasso e a decomposição em valores singulares e a termorrugulação (BRAGA *et al.*, 2019) através das RNN. Para detectar e contar o número de abelhas que entram e saem de uma colmeia, Chen *et al.* (2012) usaram o SVM (BEN-HUR *et al.*, 2002).

Observa-se que a maior parte das pesquisas apresentadas usam a temperatura dentro das colmeias como a principal métrica de monitoramento, seja monitorada sozinha ou em conjunto com outras métricas. Apesar da viabilidade, em alguns casos, de coletar umidade com o mesmo sensor de temperatura, a umidade não é tão recorrente no campo da apicultura de precisão e precisa ser mais explorada. A umidade externa à colmeia desempenha um papel fundamental

para as abelhas quando em climas quentes (ABOU-SHAARA *et al.*, 2017). As abelhas resistem melhor às condições de forte calor quando a umidade relativa está aproximadamente em 75%; abaixo de 50% a sobrevivência das abelhas é impactada negativamente, e chega em um nível crítico quando a umidade fica abaixo de 15% (ABOU-SHAARA *et al.*, 2012). A análise dos dados de peso com as vibrações das colmeias pode apresentar perspectivas promissoras e também precisa ser mais explorada. Por outro lado, o monitoramento e a interpretação de atributos de dados como áudio, vídeo e gases podem ser muito caras computacionalmente, pois os sinais de áudio e vídeo requerem equipamentos mais robustos para coleta e transmissão de dados, e os sensores de gás têm um alto custo financeiro. Especificamente para capturar o áudio, é necessário usar um procedimento sistemático para localizar os microfones dentro da colmeia, bem como para filtrar outros ruídos que não são emitidos pelas abelhas na etapa de processamento. Percebe-se também que muitos trabalhos se concentram apenas no desenvolvimento de ferramentas dos sistemas de monitoramento, ignorando a semântica intrínseca dos dados. Isso implica que a etapa de análise de dados é o principal obstáculo na abordagem da apicultura de precisão e precisa de mais atenção.

Neste trabalho, os algoritmos de aprendizado de máquina usaram dados obtidos através de sensores implantados dentro e fora das colmeias, sendo utilizados temperatura interna e peso da colmeia. Além dos sensores, o conhecimento padronizado obtido nas inspeções foi utilizado no treinamento de algoritmos para aumentar a precisão da detecção automática do bem estar das colônias. Portanto, nenhum dos artigos citados anteriormente prevê o estado geral do bem estar das colmeias, utilizando dados do ambiente em que as colônias se encontram, nem dados de inspeções realizadas pelo apicultor. Até onde se sabe, esse estudo é o primeiro a detectar e caracterizar os níveis de bem estar de uma colônia de abelhas *Apis mellifera* através do aprendizado de máquina usando os dados internos do microclima (temperatura), produtividade (peso), clima externo e dados de inspeção. A Tabela 1 resume os principais aspectos do trabalho relacionado discutidos. Na Tabela 1, os acrônimos significam: I = uso de planilhas de inspeção, IoT = uso de sensores internos, IoT + E = uso de sensores internos e externos, AM = uso do Aprendizado de Máquina, IoT + E + I = uso de sensores internos, externos e planilhas de inspeção.

Tabela 1 – Resumo dos trabalhos relacionados.

	I	IoT	IoT + E	AM	I + IoT + E + AM
(HOOPER, 1996)	✓	–	–	–	–
(BENCSIK *et al.*, 2011)	–	✓	–	–	–
(CHEN *et al.*, 2012)	–	✓	–	✓	–
(ZACEPINS; KARASHA, 2012)	–	✓	–	–	–
(STALIDZANS; BERZONIS, 2013)	–	✓	–	–	–
(KRIDI *et al.*, 2014)	–	✓	–	✓	–
(BENCSIK *et al.*, 2015)	✓	✓	–	–	–
(FITZGERALD *et al.*, 2015)	–	✓	–	–	–
(LEE *et al.*, 2015)	✓	–	–	–	–
(MURPHY *et al.*, 2015)	–	✓	–	–	–
(MURPHY *et al.*, 2015)	–	✓	–	–	–
(TASHAKKORI; GHADIRI, 2015)	–	✓	–	–	–
(CHAZETTE *et al.*, 2016)	–	✓	–	–	–
(KRIDI *et al.*, 2016)	–	✓	–	✓	–
(KVIESIS; ZACEPINS, 2016)	–	✓	✓	✓	–
(MURPHY *et al.*, 2016)	–	✓	✓	✓	–
(ZACEPINS *et al.*, 2016)	–	✓	–	–	–
(GIL-LEBRERO *et al.*, 2017)	–	✓	–	–	–
(JACOBS *et al.*, 2017)	✓	–	–	–	–
(MEIKLE *et al.*, 2017)	–	✓	–	–	–
(RUAN *et al.*, 2017)	–	✓	–	–	–
(ZACEPINS *et al.*, 2017)	–	✓	–	–	–
(CARVALHO *et al.*, 2018)	–	✓	–	✓	–
(DINEVA; ATANASOVA, 2018a)	–	✓	✓	✓	–
(SERITAN *et al.*, 2018)	–	✓	✓	✓	–
(MACIEL *et al.*, 2018a)	–	✓	✓	✓	–
(FLORES *et al.*, 2019)	✓	✓	–	–	–
(ROBLES-GUERRERO *et al.*, 2019)	–	✓	–	✓	–
Este trabalho	✓	✓	✓	✓	✓

Fonte: Elaborada pelo autor (2019)

2.4 Sumário do Capítulo

Este capítulo teve como objetivo descrever fundamentos teóricos diretamente relacionadas a esta pesquisa. Os temas abordados foram a apicultura de precisão, o aprendizado de máquina e os trabalhos relacionados. Foram apresentadas as principais características das duas pricipais abordagem de aprendizado de máquina (aprendizado não supervisionado e o supervisionado) e os algoritmos de cada abordagem que foram utilizados nesse trabalho. A partir das características descritas, foi possível identificar a melhor abordagem para cada tipo de problema relacionado com a análise de dados de sensores.

Vale destacar ainda que este capítulo teve como objetivo descrever de forma sucinta as características dos algoritmos de agrupamento e classificação. Detalhes adicionais de parâmetros

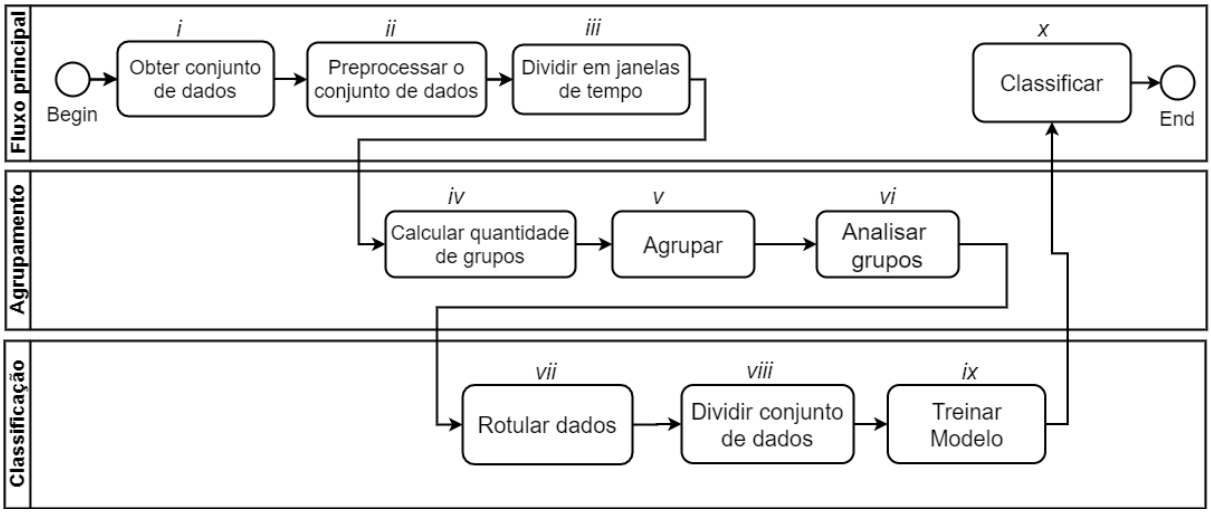
dos algoritmos serão descritos no Capítulo 3. Por fim, vale lembrar também que existem ainda diversas outras técnicas de aprendizado de máquina que não foram utilizadas nesta tese. A exemplo disso, é possível citar os processos Gaussianos (ŽGANK, 2018), a lógica Fuzzy e as regras de associação (KVIESIS *et al.*, 2020).

3 AGRUPAMENTO

Este capítulo descreve os aspectos metodológicos e os resultados da abordagem que utiliza a técnica de mineração de dados agrupamento, detalhando tópicos relacionados às ferramentas utilizadas, coleta e pré-processamento de dados, estratégias de aprendizado, bem como a análise e detecção dos níveis de bem estar de colônia de abelhas utilizando o agrupamento. Para a validação dessa abordagem, foi utilizado um conjunto de dados específico conforme recomendado no Capítulo 1.

Os níveis de bem estar das colônias foram obtidos a partir da clusterização, depois validados por especialista, para só então, acontecer a etapa de classificação de dados. Nessa abordagem, os possíveis níveis de bem estar não são definidos previamente, nem em termos quantitativos nem qualitativos, sendo a validação por um especialista uma etapa mandatória. As etapas a serem realizadas para obteção do modelo de classificação são ilustradas na Figura 9. A Seção 3.1 detalha cada uma dessas etapas. Lembrando que o processo ou método da Figura 9 é uma instância do método KDD apresentado no Capítulo2.

Figura 9 – Processo de predição via aprendizado não supervisionado



Fonte: Elaborada pelo autor (2019)

3.1 Materiais e Métodos

Para validar do processo apresentado na seção anterior, foram usados dados do portal HiveTool.net[1]. O Hivetool é um projeto de código aberto cujo objetivo é produzir ferramentas de software e hardware para monitorar e gerenciar colmeias, fornecendo monitoramento de

[1] http://www.hivetool.net/

peso, temperatura interna e ambiente, umidade interna e externa, níveis de luz, além de algumas variáveis do meio ambiente (não disponível para todas as colmeias), tais como: rajadas de vento, direção do vento, ponto de condensação da água, pressão atmosférica e chuva.

3.1.1 Conjunto de dados

O conjunto de dados utilizado possui dados de 2 colmeias dos seguintes atributos físicos: peso, temperatura e umidade relativa interna das colmeias, temperatura e umidade relativa do ambiente. Colônias de abelhas da espécie *Apis mellifera*. A Tabela 2 apresenta as colmeias utilizadas bem como a localização, período de tempo observado e a quantidade de observações (tuplas) de cada colmeia. A seguir, as variáveis utilizadas nesse trabalho são detalhadas. Para cada variável, foi tomada para a análise uma média diária, contudo, em todas as colmeias a amostragem nos sensores era feita a cada cinco minutos. A Figura 10 mostra o sistema HiveTools usado para coletar dados na colmeia Emil. Após o cálculo da média diária, procedeu-se com o pré-propcessamento. Optou-se pelo uso da média diária para tornar viável a execução do agrupamento e da classificação, utilizando-se todo o conjunto de dados, o tempo de execução dos algoritmos ficou inviável.

Tabela 2 – Sumário das colmeias, período de monitoramento e quantidade de amostras

Colmeias	Localização	Período	#amostras
Arnas	Rebild, Jutlândia, DK	01/03/2017 to 28/02/2018	92.749
Emil	Grimstad, Agder, NO	01/03/2017 to 28/02/2018	88.181

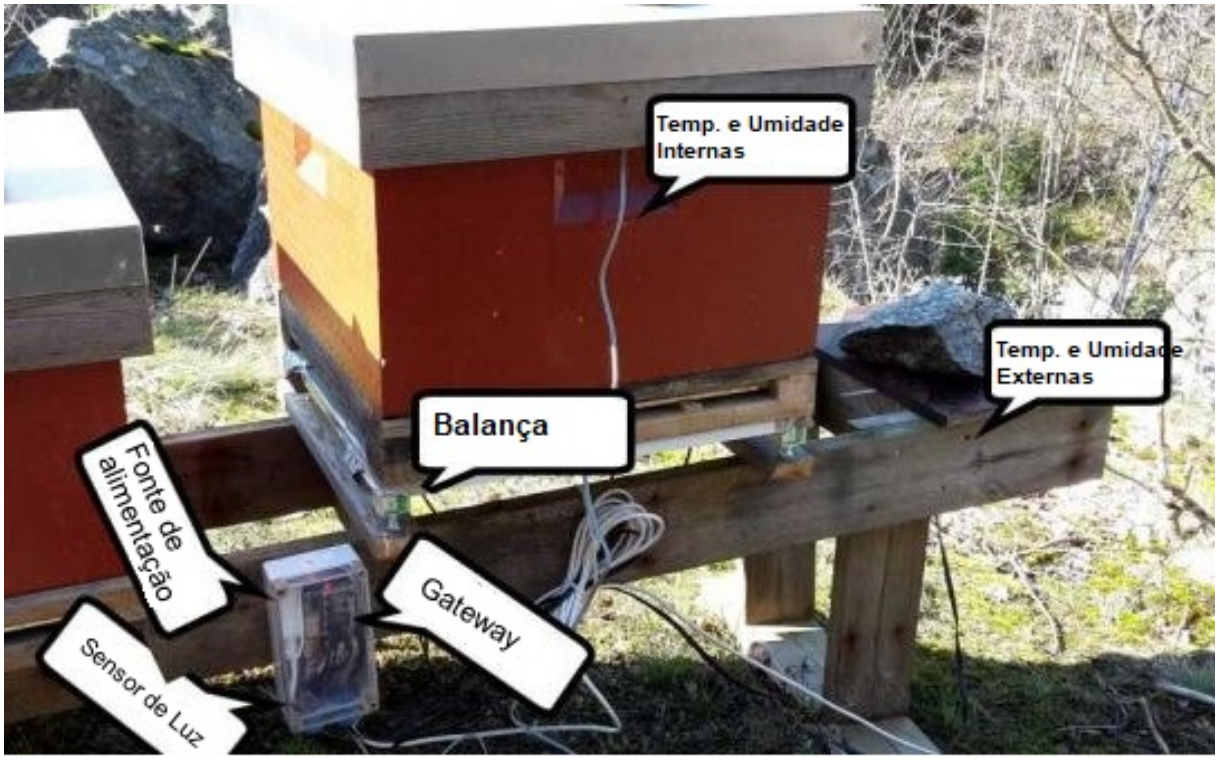
Fonte: Elaborada pelo autor (2019)

1. Temperatura interna em °Celsius: temperatura obtida com um sensor colocado no centro da colônia. Essa variável é uma das mais importantes, pois indica a termorregulação da colônia;
2. Umidade interna em %: obtido com um sensor colocado no centro da colônia. Esse recurso, juntamente com a temperatura, indica a termorregulação da colônia e outros fenômenos importantes da colônia, tais como início da produção de uma nova cria e a preparação para o inverno;
3. Peso da colmeia em kg: peso da caixa de madeira mais o peso da colônia obtido por meio de uma balança digital. Essa variável indica a produtividade da colônia e, consequentemente, a quantidade de reserva alimentar, podendo indicar indiretamente também o nível de bem

estar da colônia;

4. Variação do peso em kg: um valor que indica se houve variações (positivas ou negativas) no peso da colmeia calculado para cada período de 7 dias. Esta variável pode indicar se a colônia está conseguindo acumular reservas de alimento ou em que taxa está consumindo as reservas que já possui.

Figura 10 – Sistema de monitoramento HiveTool na colmeia Emil.



Fonte: http://www.evit.no/wp-content/uploads/2016/05/emil-scale2.jpg

3.1.2 Preprocessamento

A atividade de pré-processamento de dados é composta principalmente por duas etapas: remoção de anomalias e redimencionamento dos dados.

3.1.2.1 Detecção e remoção de anomalias

Anomalias são dados cujos valores são muito diferentes dos outros ou estão fora dos intervalos aceitáveis do conjunto de dados. Anomalias podem influenciar o resultado e consequentemente causar distorções. Para o método aplicado baseado em agrupamento, a detecção e remoção da anomalia foi realizada pelo método Tukey (TUKEY, 1977). O método Tukey define um valor discrepante como aqueles valores no conjunto de dados que estão

distantes do ponto médio (mediana). A distância máxima ao centro dos dados é chamada de parâmetro de limpeza. Caso os dados estejam fora desse intervalo, eles são considerados uma anomalia (*outlier*). Se por acaso o parâmetro de limpeza for muito grande, o teste ficará menos sensível a anormalidades. Alternativamente, se for muito pequeno, muitos valores são detectados como outliers. Esse método funciona bem quando existem valores extremos e eles podem ser facilmente detectados. Os limites são calculados pela Equação 3.1, em que $Q1$ e $Q3$, que são, respectivamente, o primeiro e o terceiro quartis de um atributo do conjunto de dados.

$$[Q_1 - 1.5 \times (Q_3 - Q_1), Q_3 + 1.5 \times (Q_3 - Q_1)]. \tag{3.1}$$

3.1.2.2 Padronização de dados

Após remover os valores discrepantes, prosseguimos para a etapa de normalização, necessária para que os dados estejam na mesma escala nas próximas etapas da análise. Os algoritmos de agrupamento e classificação são sensíveis à escala dos dados de entrada, os dados foram redimensionados para serem usados nos algoritmos. Esse processo também é geralmente usado para melhorar a estabilidade numérica de um modelo de predição. O redimensionamento de dados lida com parâmetros de diferentes unidades e escalas, especialmente quando se deve comparar valores e, para isso, eles precisam ter a mesma escala para obter resultados positivos. Assim, cada atributo "x" do conjunto de dados tem um valor normalizado "x_{new}" padronizado através da transformação pela média e variância ou transformação *z-score* (Eq. 3.2) (KREYSZIG, 2010).

A padronização transforma os dados para ter uma média (μ) igual a 0 e um desvio padrão (σ) igual a 1 (variação unitária). Para padronizar os dados, a centralização é feita de tal forma que o valor médio do preditor é subtraído de todos os valores. Como resultado dessa centralização, o preditor tem uma média zero. Também fazemos o agendamento no qual cada valor da variável preditora é dividido por seu desvio padrão. Assim, a escala dos dados força os valores a terem um desvio padrão comum de 1.

$$x_{new} = \frac{x - \mu}{\sigma}. \tag{3.2}$$

3.1.3 Dividir em janelas de tempo

De maneira geral, o ciclo anual de colônias de abelhas em climas temperados pode ser dividido em dois períodos: a época das estações mais frias do ano (outono e inverno), quando as abelhas ficam menos ativas; e a época das estações mais quentes do ano (primavera e verão), quando há alta atividade das abelhas (KVIESIS; ZACEPINS, 2016). Durante esses períodos, diversos estágios do ciclo de vida normal das colônias podem ser observados. A divisão proposta, visa caracterizar esses estágios normais, para obter os valores típicos das grandezas sensoriadas nas quatro estações do ano. Para o apicultor, ter o conhecimento sobre qual estado uma colônia está em determinado momento, sem abrir a colmeia, possibilita administrar melhor seu apiário, evitando perdas e maximizando a produtividade (KVIESIS; ZACEPINS, 2016).

Assim, será possível caracterizar padrões para intervalos de tempo menores, e revelar o comportamento cíclico das abelhas. Os conjuntos de dados descritos na Seção 3.1.1 foram divididos em dois períodos de seis meses: (*i*) de março a agosto e (*ii*) de setembro a fevereiro. O primeiro período corresponde às estações primavera e verão (período ativo), e o segunda corresponde às estações outono e inverno no hemisfério norte (período de menos atividades) (KVIESIS; ZACEPINS, 2016).

3.1.4 Calcular quantidade de grupos

A metodologia de aplicação do agrupamento utilizada neste trabalho baseia-se na metodologia apresentada por (HAN *et al.*, 2011), com a inclusão da etapa de cálculo da quantidade ótima de grupos definida pelo através do índice CH, conforme descrito na Seção 2.2.1.1. As etapas que envolvem o cálculo da quantidade de grupos e do agrupamento em si são apresentadas no Algoritmo 1.

A primeira etapa consiste em calcular o índice CH para determinar o melhor número (k_{opt}) de partes que pode dividir o conjunto de dados para cada período de 6 meses. O número máximo de grupos (k_{max}) representa o número máximo de estágios de uma colmeia. Inicialmente, o valor de k_{max} foi definido como $k_{max} = 24$ para executar uma análise exploratória para valores de k entre 2 e 24. Se os valores mais altos do índice CH ocorrerem para $k >= 20$ em qualquer período, uma exploração será feita para $k > 24$. Caso o índice de CH mais alto ocorra para $k = 24$, haveria uma média de 1 grupo por semana.

Para cada valor de K (etapa 1.1), o algoritmo k-means é aplicado por N_r rodadas

Algoritmo 1: Agrupamento com *k-means*

Data: Conjunto de dados (sem anomalias e padronizado), K_{max} e N_r

Result: Grupos

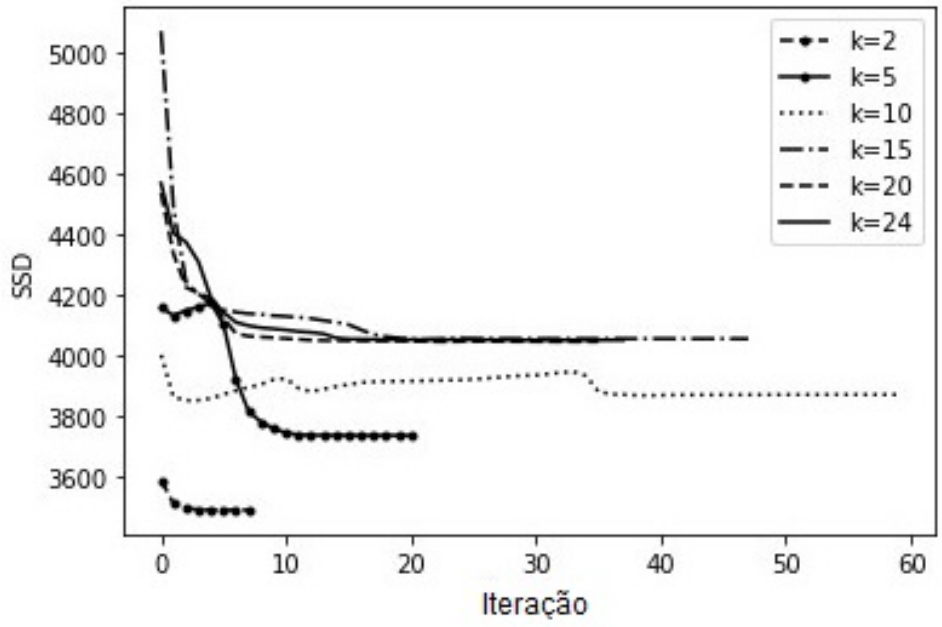
1. Calcular o índice CH para um conjunto de K grupos.
   $K = \{2, \ldots, k_{max}\}$
   for each $k \in K$ **do**
   1.1 Aplique o algoritmo k-means por N_r rodadas.
   1.2 Escolha os protótipos de rodadas com menor *SSD*.
   1.3 Calcule o valor do índice CH para k.
   endfor
2. Escolha o valor ótimo, k_{opt}, que otimiza o índice CH.
3. Particione o conjunto de dados em k_{opt} grupos usando a Distância Euclidiana.
4. Reporte estatísticas descritivas dos atributos por agrupamento.

até sua convergência (até que a posição do protótipo w_i (Eq. 2.4) pare de mudar). Como os centróides iniciais são escolhidos aleatoriamente, sua posição final pode variar a cada execução do algoritmo. Portanto, o número de rodadas (N_r) deve ser suficiente para obter os centróides que possuem o menor *SSD*. Para garantir a convergência, definiu-se 300 iterações ($N_r = 300$) para calcular o custo de todos os pontos. A tolerância relativa em relação ao *SSD* para declarar convergência foi de 0,001.

Para a escolha dos melhores protótipos (etapa 1.2), o *SSD* foi usado como critério de decisão. Um valor menor de *SSD* significa distâncias menores entre os dados dos grupos e seus respectivos centróides. Na Figura 11 é possível observar a evolução do valor de *SSD* para $k = 2, 5, 10, 15, 20$ e 24 para o 1° período no conjunto de dados Arnas. Neste exemplo, a convergência de w_i (Eq. 2.4) ocorreu na iteração 9 para $k = 2$. Esse comportamento em relação à convergência do k-means se repete nos demais períodos. Em seguida, o índice CH foi calculado para determinar a quantidade de partições mais apropriada para cada período (etapa 1.3).

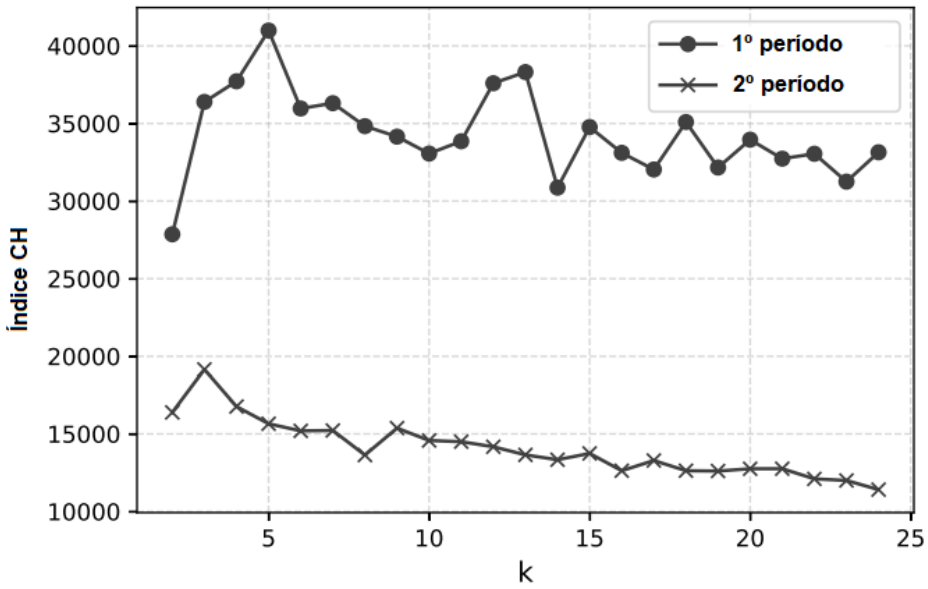
Com os melhores protótipos de cada valor de k definidos, eles foram usados para calcular o índice de CH. O valor ideal de k (k_{opt}) é aquele que apresentou o índice CH mais alto (otimiza o índice CH) (etapa 2). A Figura 12 mostra o valor do índice CH para cada valor de k calculado para o conjunto de dados Arnas. Para o primeiro período, o valor mais alto do índice ocorre para $k = 5$ e no segundo período para $k = 3$, sendo estes os números apropriados de grupos para o conjunto de dados Arnas. A Figura 13 mostra o valor do índice CH para cada valor de k calculado para o conjunto de dados Emil. No primeiro período, o valor mais alto do índice ocorre para $k = 3$ e no segundo período para $k = 6$, sendo estes as quantidades apropriadas de grupos para o conjunto de dados Emil.

Figura 11 – Evolução dos valores de *SSD* para k = 2, 5, 10, 15, 20 e 24.



Fonte: Elaborada pelo autor (2019)

Figura 12 – Índice CH para $2 <= k <= 24$ no conjunto de dados Arnas.

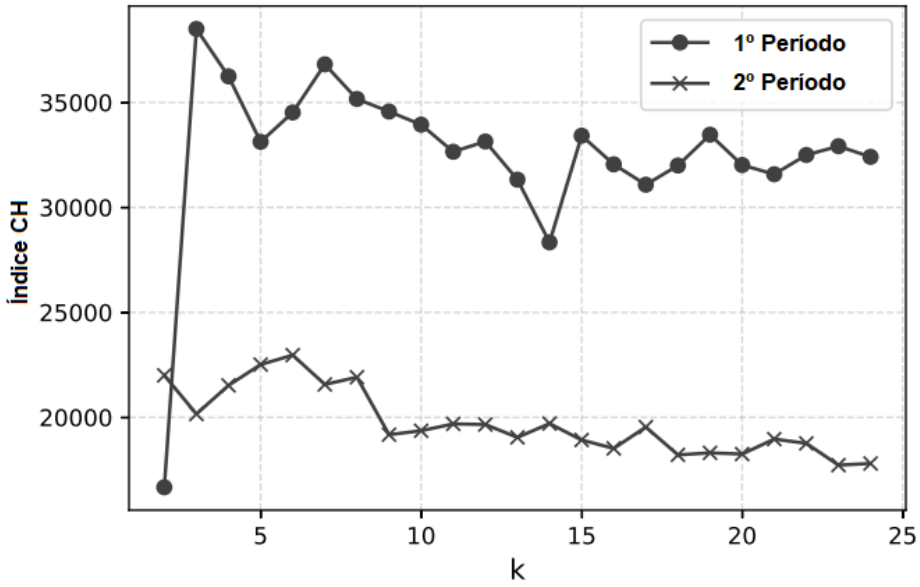


Fonte: Elaborada pelo autor (2019)

3.1.5 Agrupar

Após definir o número ideal de grupos para cada período, prosseguiu-se com o particionamento do conjunto de dados normalizados em k_{opt} grupos com a execução do *k-means* usando o critério da Distância Euclidiana para o protótipo mais próximo (etapa 3). Em seguida, fez-se a obtenção das estatísticas descritivas (etapa 4) dos atributos de cada agrupamento para melhor análise dos grupos pelo especialista.

Figura 13 – Índice CH para 2 <= k <= 24 no conjunto de dados Emil.



Fonte: Elaborada pelo autor (2019)

3.1.6 Analisar grupos

Os grupos obtidos na fase de agrupamento, são, então, apresentados a um especialista em apicultura para interpretação. O especialista leva em consideração: as estatísticas descritivas (que inclui média, mediana, moda, desvio padrão, variância, valor máximo e mínimo, assimetria e curtose), o período do ano (estação), a região geográfica, a espécie da abelha em estudo e a quantidade de colmeias.

O objetivo dassa fase é buscar a semântica de cada grupo, haja vista que o agrupamento feito pelo *k-means* leva em conta apenas a estrutura espacial dos dados. Essa fase é necessária para cada nova execução do algoritmo de agrupamento ou quando novas colmeias forem adicionadas ao conjunto de dados, da mesma espécie de abelhas ou de uma espécie diferente.

3.1.7 Rotular dados

Após a associação dos grupos com níveis de bem estar ou do ciclo de vida das colônias de abelhas, cada ponto amostral recebeu um rótulo referente ao agrupamento do qual faz parte. Esses rótulos foram aplicados aos pontos amostrais de dados não normalizados, para que a correspondência entre as posições de cada ponto amostral nos dois conjuntos de dados fosse obedecida.

Após a rótulagem de cada ponto amostral do conjunto de dados, as três técnicas de classificação apresentadas na Seção 2.2.2 foram aplicadas com o objetivo de definir o algoritmo de classificação mais preciso para a criação de um modelo de classificação que possa ser usado para categorizar novos pontos amostrais obtidos através dos sensores.

3.1.8 Dividir conjunto de dados

Para realizar os experimentos com os algoritmos de classificação, o conjunto de dados rótulado foi aleatoriamente separados nos conjuntos de treinamento, teste e validação com a seguinte proporção: 60% do conjunto de dados para treinamento, 20% para teste e 20% para validação. Uma técnica de validação cruzada, do inglês *Cross-Validation* (CV) foi usada no conjunto de treinamento para impedir que uma única porção do conjunto de dados fosse usada. O CV foi configurado com 10 camadas (KOHAVI, 1995).

3.1.9 Otimizar hiperparâmetros

Para o algoritmo kNN, o parâmetro de avaliação utilizado foi a precisão do método para diferentes valores de k e p, onde p é a ordem da distância de Minkowski. A distância de Minkowski é calculada de acordo com a Equação 3.3. No caso em que $p = 1$, a distância é equivalente à distância de Manhattan e, no caso em que $p = 2$, a distância é equivalente à distância euclidiana. Foram avaliados $p = 1$ e $p = 2$.

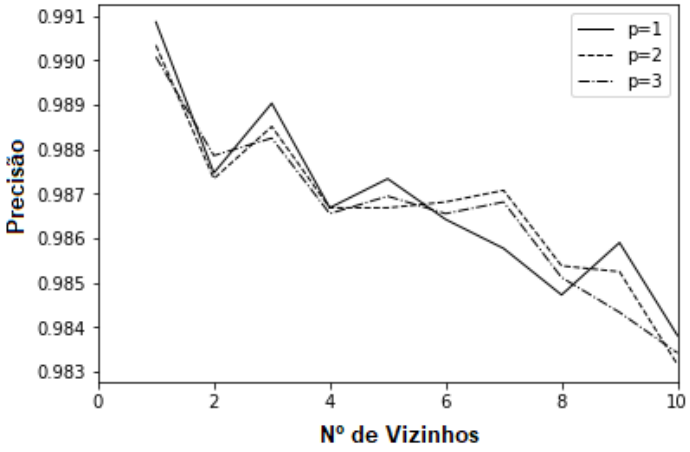
$$d_{XY} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{3.3}$$

Em relação a k, um vetor com valores de 50 de k foi definido no intervalo $[1,50]$. Todas as combinações possíveis de k e p foram testadas. Como pode ser visto na Figura 14, a maior precisão média ocorre para $k = 1$ e $p = 1$.

Para o algoritmo *Random Forest*, os hiperparâmetros avaliados foram o número de árvores ($trees$) e o número de características utilizadas na criação de cada árvore ($mtry$). Para o $mtry$, foram testados os valores no intervalo $[1,6]$. Para $trees$, de acordo com (OSHIRO *et al.*, 2012), o número inicial de árvores para cada característica n está entre 64 e 128. Para obter o valor ideal das árvores no conjunto de validação uma verificação cruzada 10 dobras foi executada.

Para o algoritmo *Naive Bayes*, como os pontos amostrais do conjunto de dados são de natureza contínua, o kernel Gaussiano foi usado. Portanto, não havia hiperparâmetros a

Figura 14 – Precisão de k (de 1 a 10) vs. p.



Fonte: Elaborada pelo autor (2019)

serem avaliados para encontrar a combinação da melhor precisão média. Assim, a precisão foi calculada a partir de 50 repetições, das quais foram extraídas a precisão média e o desvio padrão.

3.2 Resultados e Discussões

Os centróides dos grupos obtidos nos experimentos podem ser vistos na Tabela 3, em que "T", "RH" e "W", respectivamente, referem-se à temperatura, umidade relativa e peso das colmeias.

Tabela 3 – Centróides do grupos

	Arnas						Emil					
	1° período			2° período			1° período			2° período		
	T(°C)	RH(%)	W(Kg)	T(°C)	RH(%)	W(Kg)	T(°C)	RH(%)	W(Kg)	T(°C)	RH(%)	W(Kg)
C0	14.3	89.2	25.3	10.3	78.1	31.0	26.5	55.0	12.4	13.5	79.9	15.6
C1	29.9	53.3	23.3	13.4	93.3	31.2	29.4	69.1	14.3	7.2	82.1	14.5
C2	30.8	54.0	11.1	7.7	96.0	29.5	10.9	75.8	14.0	7.0	80.6	16.5
C3	33.6	58.4	37.3							5.9	74.4	14.8
C4	16.4	85.9	24.6							7.9	78.4	15.2
C5										16.3	74.2	16.9

Fonte: Elaborada pelo autor (2019)

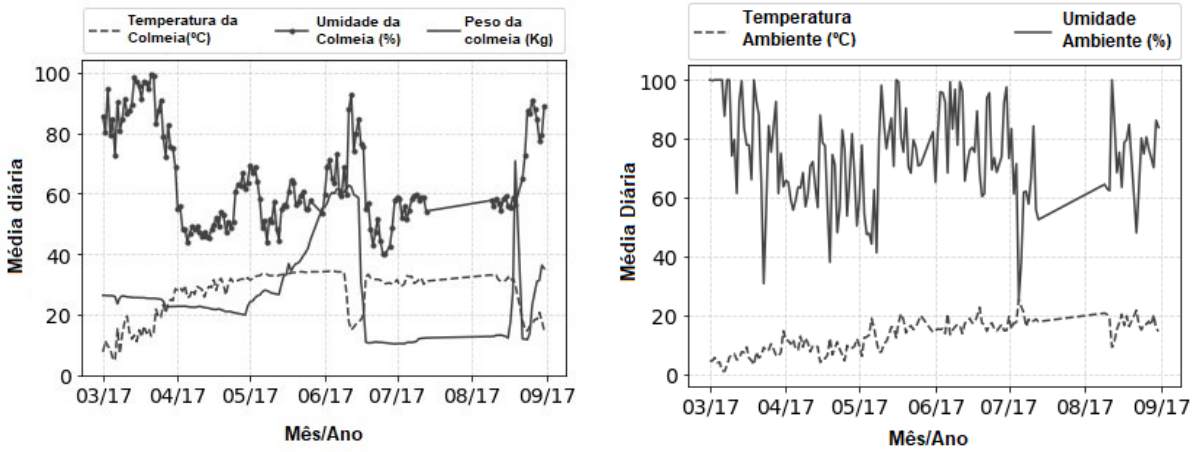
A interpretação e validação dos grupos obtidos foi realizada por um especialista em apicultura e pelo conhecimento disponível em artigos sobre o assunto. No primeiro período, de Março a Agosto de 2017, a metodologia utilizada retornou 5 grupos para colmeia do apiário de Arnas. Seus centróides são mostrados na Tabela 3, nas colunas sob a descrição "Arnas 1° período".

Dois grupos têm centróides semelhantes: o Grupo 0 (C0) e o Grupo 4 (C4), com 13.519 e 3.690 pontos amostrais, respectivamente. Eles estão associados ao estado da colônia

durante a transição entre as estações frias e quentes do ano (do inverno para a primavera e do verão para o outono, respectivamente). Essa conclusão é alcançada pelos valores de temperatura (o mais baixo) e umidade relativa (o mais alto) (MEIKLE *et al.*, 2017). Isso se deve à influência do período mais frio e mais úmido nas colônias, veja a Figura 15(b). Essa alta temperatura, influenciada principalmente pela temperatura externa, foi discutida em mais detalhes por Rice (2013) e Stalidzans e Berzonis (2013). No caso do Grupo 0 (C0), do inverno para a primavera, as abelhas geralmente **iniciam a produção de crias**, pois durante esse período é possível observar também a faixa ideal de umidade interna para a incubação normal dos ovos, variando de 90% a 95% (ABOU-SHAARA *et al.*, 2017). A Figura 15(a) mostra a média diária de temperatura, umidade e peso da colmeia Arnas durante o primeiro período estudado. Observe que no final de março a umidade no interior da colmeia cai drasticamente, por causa do final do inverno.

O Grupo 4 (C4), possui uma quantidade menor de pontos amostrais, uma temperatura mais alta e uma umidade mais baixa que o grupo 0, está associada ao final do verão/início do outono. Período do ano em que as abelhas começam a se preparar para temperaturas externas progressivamente mais baixas ou um período de **preparação para a passagem pelo inverno**.

Figura 15 – Média diária dos atributos para a colônia Arnas no 1° período.



(a) Atributos internos da colônia.

(b) Atributos externos da colônia.

Fonte: Elaboradas pelo autor (2019)

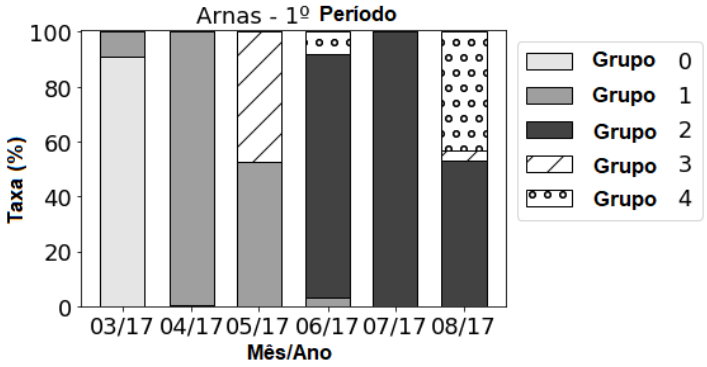
O Grupo 1 (C1), com 8.022 pontos amostrais, está associado ao estado da colônia no meio da primavera. Isso é evidenciado principalmente pela baixa umidade relativa, devido a baixa incidência de precipitações nesta estação e, consequentemente, a umidade relativa apresenta médias baixas, como pode ser visto na Figura 15(b). Sob essas condições, as abelhas geralmente começam a evaporar a água do néctar e forragear para realizar a coleta de água

para manter a RH alta durante a primavera (ABOU-SHAARA *et al.*, 2017), como pode ser visto na Figura 15(a). Além disso, intensificam o armazenamento de pólen, o número de abelhas forrageadoras, a atividade **produção de cria** e a **coleta e armazenamento de alimentos** (STALIDZANS; BERZONIS, 2013; NEUPANE; THAPA, 2005).

O Grupo 2 (C2), com 10.936 pontos amostrais, é interpretado como o estado da colônia no verão, não apenas pela baixa umidade relativa (semelhante à observada em C1), mas também pelo baixo peso. O baixo valor do peso pode indicar a **extração de mel** produzido pelas abelhas durante a primavera pelo apicultor ou pode estar associado ao **enxameamento**; nesse caso específico, o enxameamento é menos provável, devido à diferença significativa de peso (SEELEY; VISSCHER, 1985). O Grupo 3 (C3), com 2.535, é entendido como o estado da colônia no final da primavera. Nesse momento, ocorre o **ápice da produção de mel**, o que se reflete no alto valor do peso do centróide (SEELEY; VISSCHER, 1985).

Esses resultados podem ser verificados na Figura 16, que mostra a proporção do número de pontos amostrais de cada grupo durante os meses do primeiro período em ordem temporal. Note-se que os pontos amostrais de C0 são distribuídos principalmente no mês de março e os pontos amostrais de C4 no mês de agosto. Estes meses correspondem, respectivamente, ao início da primavera (depois do inverno) e ao final do verão (antes do outono). Os pontos amostrais de C1 são distribuídos principalmente entre os meses de março, abril e maio, que correspondem à primavera. Os pontos amostrais de C2 são distribuídos entre os meses de junho, julho e agosto, que correspondem ao verão. Finalmente, os pontos amostrais C3 são distribuídos principalmente no mês de maio, o último mês da primavera no hemisfério norte.

Figura 16 – Distribuição dos grupos nos meses para a colmeia Arnas no 1° período.



Fonte: Elaborada pelo autor (2019)

No segundo período, de setembro/2017 a fevereiro/2018, a metodologia utilizada retornou como resultado 3 grupos para a colmeia do apiário de Arnas, cujos centróides são

mostrados na Tabela 3, nas colunas sob a descrição “Arnas 2° period”.

Por terem as temperaturas mais altas, o Grupo 0 (C0) e o Grupo 1 (C1), com 10.917 e 7.315 pontos amostrais respectivamente, são entendidos como o estado da colônia durante a transição entre as estações (de outono ao inverno e inverno à primavera) (BRAGA *et al.*, 2019).

Como a temperatura e a umidade em C0 são menores que em C1, pode-se supor que C0 esteja relacionado à transição do inverno para a primavera. Nesse período, é possível destacar que a umidade dentro da colmeia é menor que a umidade externa Figura 17(a) e Figura 17(b). Isso está relacionado à capacidade de **termorregulação** da colmeia, a fim de manter a temperatura dentro da colmeia mais alta que a temperatura externa (BRAGA *et al.*, 2019).

Como C1 tem a temperatura mais alta e umidade relativa alta também se comparado com C0, é possível supor que se refira à transição do outono para o inverno, pois dentro da colmeia a temperatura média é mais alta no outono, como é possível ver na Figura 17(a).

Complementarmente, C2, com 16.441 pontos amostrais, é interpretado como o estado da colônia no meio do segundo período, quando começa a **passagem pelo inverno**. Neste momento, a temperatura mais baixa e a umidade relativa mais alta ocorrem, ver Figura 17(b).

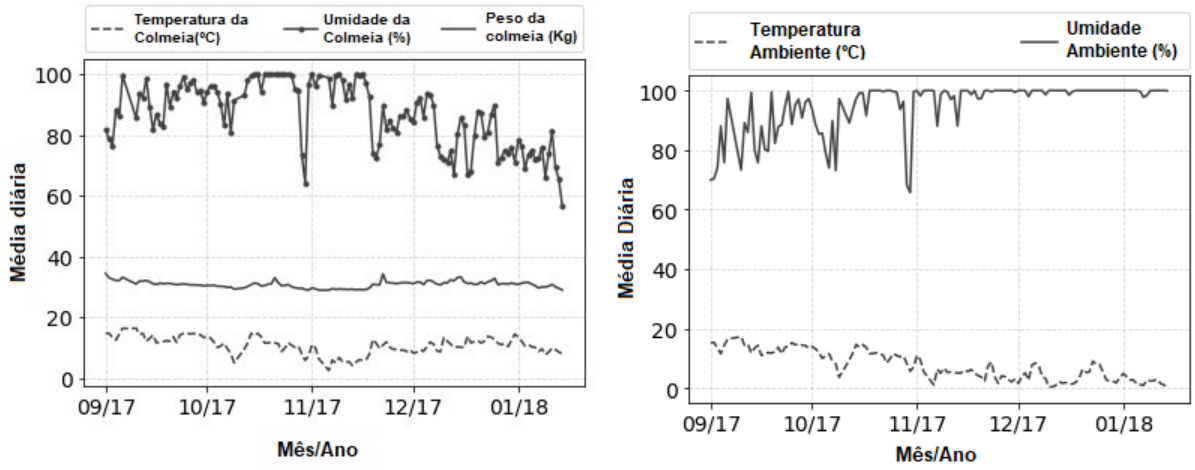
É importante observar que, se a temperatura da colônia (Figura 17(a)) seguisse a tendência da temperatura ambiente observada na Figura 17(b), o valor da temperatura C0 deveria ser menor. No entanto, é maior que o valor da temperatura de C2. Isso indica a capacidade de **termorregulação** da colônia para se **prepara para a passagem pelo inverno**, evidenciada também pelo valor alto e regular do peso (reservas de mel), pois o controle térmico depende da presença de um número maior de abelhas trabalhadoras saudáveis e das reservas de mel (COOK; BREED, 2013).

Esses resultados podem ser verificados pela Figura 18, que mostra a proporção do número de pontos amostrais de cada grupo durante os meses do segundo período em uma ordem temporal. Pode-se observar que os pontos amostrais de CO são distribuídos principalmente nos últimos meses do segundo período, o que corresponde ao final do inverno (antes da primavera). Os pontos amostrais C1 são distribuídos principalmente em setembro e outubro, que correspondem ao início do outono. Finalmente, os pontos amostrais de C2 são distribuídos principalmente entre outubro e novembro, quando o outono termina e o inverno começa.

Para a colmeia do apiário Emil, a metodologia retornou como resultado 3 grupos no primeiro período. Seus centróides são mostrados na Tabela 3.

Semelhante à avaliação feita para a colmeia de Arnas, o C2, com 14.317 pontos

Figura 17 – Média diária dos atributos para a colmeia Arnas no 2° período.

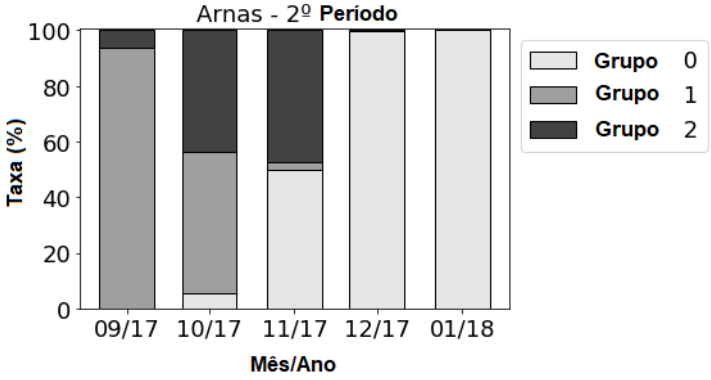


(a) Atributos internos da colmeia.

(b) Atributos externos da colmeia.

Fonte: Elaboradas pelo autor (2019)

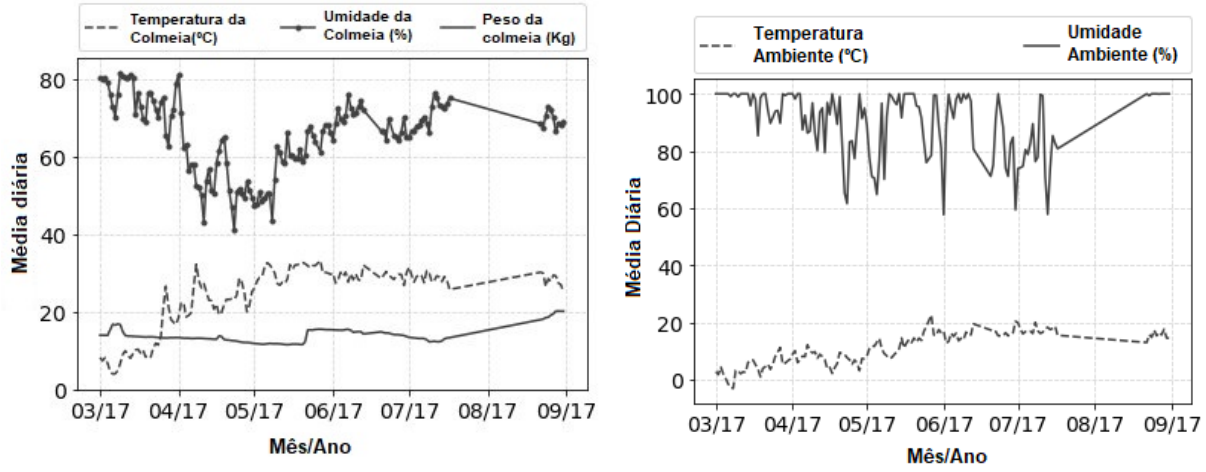
Figura 18 – Distribuição dos grupos nos meses para a colmeia Arnas no 2° período.



Fonte: Elaborada pelo autor (2019)

amostrais, é interpretado como o estado da colônia durante a transição entre as estações frias e quentes do ano, do inverno para a primavera, devido ao menor valor de temperatura e maior UR valor. Nesse período, as abelhas geralmente **começam a produção de cria**. Isso se deve à influência do período mais úmido nas colônias entre março e abril, veja a Figura 19(b). No mesmo período, como em Arnas, a umidade no interior da colmeia diminuiu drasticamente quando o inverno terminou, veja a Figura 19(a). Ainda levando em consideração o valor da temperatura, C0 é interpretado, com 8.779 pontos amostrais, como o estado da colônia na primavera, quando geralmente ocorre a intensificação da atividade de **produção de cria**, **coleta e armazenamento de alimentos** . C1, com 13.364 pontos amostrais, é interpretado como o estado da colônia no verão, quando geralmente ocorre o **ápice da produção de mel** e a **preparação para o inverno**, uma vez que possui o maior valor de temperatura e um valor mais alto de RH, devido à transição para o outono, veja a Figura 19(b).

Figura 19 – Média diária dos atributos para a colmeia Emil no 1° período.



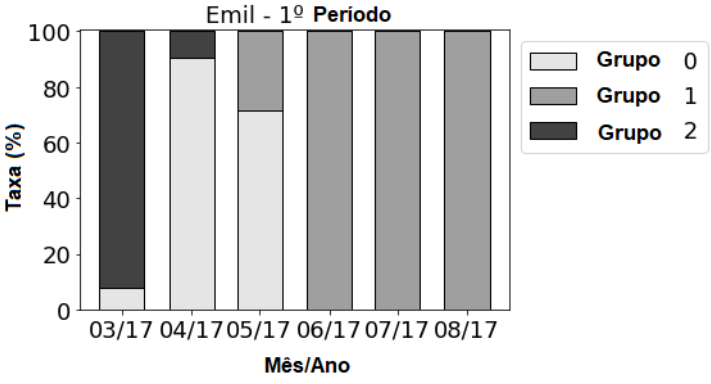
(a) Atributos internos da colmeia.

(b) Atributos externos da colmeia.

Fonte: Elaboradas pelo autor (2019)

Esses resultados podem ser verificados pela Figura 20, que mostra a proporção do número de pontos amostrais de cada grupo durante os meses do primeiro período em uma ordem temporal. Percebe-se que os pontos amostrais do C2 estão distribuídas principalmente no mês de março, primeiro mês após o inverno no hemisfério norte. Os pontos amostrais de C0 são distribuídos principalmente nos meses de abril e maio, que correspondem ao final da primavera. Por fim, os pontos amostrais de C1 são distribuídos principalmente nos meses de junho, julho e agosto, que correspondem ao verão.

Figura 20 – Distribuição dos grupos nos meses para a colmeia Emil no 1° período.



Fonte: Elaborada pelo autor (2019)

No segundo período, a metodologia apresentada retornou como resultado 6 grupos para a colmeia do apiário Emil, cujos centróides são apresentados na Tabela 3.

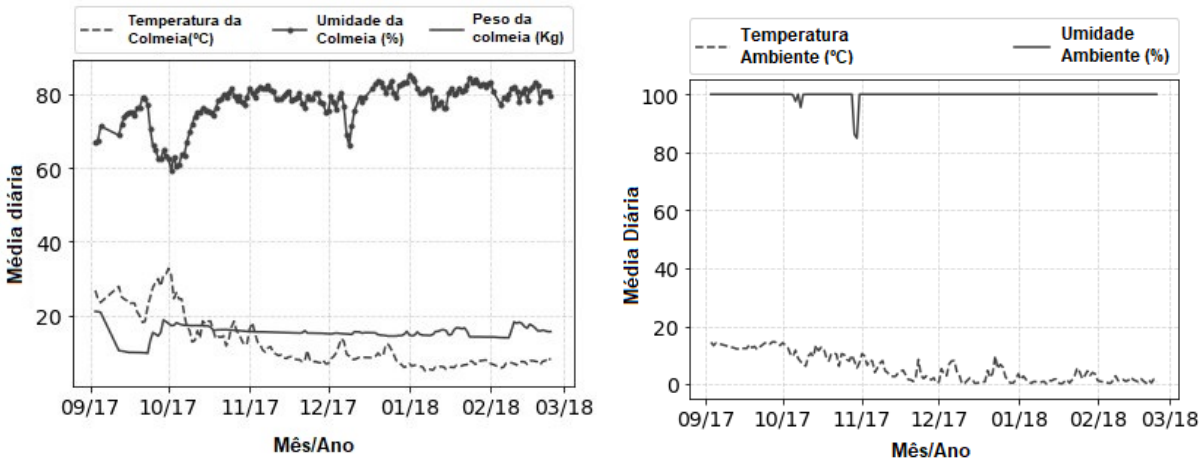
Dois grupos têm centróides semelhantes: os grupos C5 e C0, com 3,124 e 9,116 pontos amostrais, respectivamente. Eles são entendidos como o estado da colônia no início do

inverno, quando a temperatura ainda não é muito baixa. Como pode ser visto na Figura 21(a), a temperatura interna na colmeia começa a diminuir em setembro/2017, sobe repentinamente no início de outubro para promover **termorregulação** devido à súbita redução da RH. No entanto, há uma tendência de queda até o início de janeiro (grupos C4 e C3), com 5.384 e 9.499 pontos amostrais, respectivamente. A partir de janeiro é possível observar um pequeno aumento de temperatura (grupos C2 e C1), com 7.218 e 5.270 pontos amostrais respectivamente, **passagem pelo inverno**.

Na Figura, 22 é possível perceber que a proporção de grupos distribuídos nos meses do segundo período segue a ordem temporal da temperatura, que corresponde à tendência da temperatura ambiente, mostrada na Figura 21(b). Em setembro/2017, primeiro mês após o verão, existem pontos amostrais apenas para C5, com 3,124 pontos amostrais, com o maior valor de temperatura. Já em fevereiro de 2018, predominam os pontos amostrais de C2 e C3, que apresentam os menores valores de temperatura, a **passagem pelo inverno**.

É possível notar que o valor do peso dos centróides é menor em comparação ao Arnas nos dois períodos, exceto em um grupo, isso indica que o controle térmico da colmeia do apiário Emil está comprometido. Isso é visível principalmente no segundo período, que corresponde ao outono e inverno. Meikle *et al.* (2017) observou que a falta de controle da temperatura interna está diretamente associada ao bem estar da colmeia. Os autores observaram que, antes e durante o inverno, colmeias utilizadas em polinização comercial tem maior probabilidade de ficar doentes e com controle de temperatura reduzido.

Figura 21 – Média diária dos atributos para a colmeia Emil no 2° período.

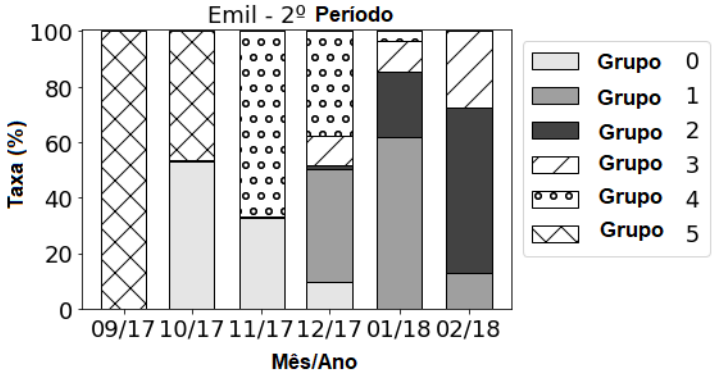


(a) Atributos internos da colmeia.

(b) Atributos externos da colmeia.

Fonte: Elaboradas pelo autor (2019)

Figura 22 – Distribuição dos grupos nos meses para a colmeia Emil no 2° período.



Fonte: Elaborada pelo autor (2019)

Após a validação e análise dos grupos, um resumo dos estados identificados (classes) é apresentado abaixo.

1. **Começo da produção de crias** - Nesse estado, a temperatura interna da colmeia aumenta como resultado da chegada da primavera e a RH ainda está alta (entre 90% e 95%), então, as abelhas começam a colocar ovos para a produção de abelhas operárias.
2. **Produção de crias, coleta e armazenamento de alimentos** - Durante esse estado, as abelhas mantêm a temperatura da cria entre 33 e 36 graus Celsius. Intensificando a produção de crias, para isso, as abelhas também precisam trabalhar na coleta e armazenamento de alimentos, as abelhas tentam manter uma RH alta dentro da colmeia, em torno de 55%.
3. **Ápice da produção de mel** - Após a coleta e armazenamento dos alimentos, ocorre o ápice da produção de mel. Esse estado é caracterizado pelo alto valor do peso da colmeia, a temperatura de aproximadamente 33 graus Celsius e a umidade relativa entre 50-60%.
4. **Extração de mel** - O apicultor faz a colheita do mel. Após a colheita, é observada uma redução repentina no peso da colmeia.
5. **Preparação para a passagem pelo inverno** - As abelhas começam a preparação para o período frio com mais acúmulo de mel e produção de abelhas operárias.
6. **Passagem pelo inverno** - Nesse estado, a temperatura mais baixa e a umidade relativa mais alta ocorrem dentro da colmeia. O peso da colmeia tem pouca variação e pode indicar o nível de bem estar da colmeia, já que a quantidade de alimentos armazenados e de abelhas operárias pode comprometer a termorregulação.

Após obter os grupos e rótulos de cada pontos amostral do conjunto de dados, os algoritmos NB, RF e kNN apresentados na Seção 2.2.2 foram aplicados com o objetivo de definir o algoritmo de classificação mais preciso para criar um modelo que fosse usado para categorizar

novos pontos amostrais. Com base nas matrizes de confusão, calculou-se a taxa de precisão de cada algoritmo. A Tabela 4 mostra que o algoritmo RF foi o classificador mais preciso, com uma precisão média de 98,77%.

Tabela 4 – Precisão dos algoritmos de classificação.

Algoritmo	**Precisão**				**Média**
	Arnas		**Emil**		
	1o Período	**2o Período**	**1o Período**	**2o Período**	
NB	93.71 ± 0.31	88.41 ± 0.38	96.76 ± 0.19	92.04 ± 0.29	92,73%
kNN	98.96 ± 0.50	96.53 ± 1.18	99.63 ± 0.65	99.30 ± 0.56	98,60%
RF	99.13 ± 0.56	97.07 ± 1.08	99.67 ± 0.41	99.24 ± 0.86	98.77%

Fonte: Elaborada pelo autor (2019)

3.3 Sumário do Capítulo

Apesar das altas taxas de acerto obtidas nos algoritmos de classificação, é importante destacar algumas precauções que se deve ter ao aplicar o método baseado em agrupamento aqui apresentado. Primeiro, vale destacar que a análise dos grupos obtidos pelo agrupamento por um especialista é mandatória. Pois, a geração dos grupos pelo *k-means* é feita apenas levando em conta a distribuição do dados no espaço vetorial. Assim, é provável a obtenção de grupos sem correlação com o ciclo de vida das colônias de abelhas. Nesse caso, deve-se ter a precaução de se realizar a rotulagem dos pontos amostrais para a etapa de classificação apenas de acordo com os grupos que possuam correlação e semântica com o ciclo de vida das abelhas ou fenômeno de interesse em estudo.

A quantidade grupos também é uma outra etapa que requer precauções. Mais uma vez, no método sugerido (índice CH) a quantidade de grupos é definida levando em conta apenas a distribuição espacial do conjunto do dados. Contudo, essa quantidade poderia inclusive ser definida arbitrariamente a fim de obter determinada quantidade grupos que se sabe previamente que ocorre no conjunto de dados. Outro parâmetro que pode ser livremente ajutado de acordo com o fenômeno de estudo é a janela de tempo. Para a validação do método foi selecionada uma jenela de 6 meses, contudo, janelas de 1 semana ou mesmo de 1 ano são igualmente possíveis.

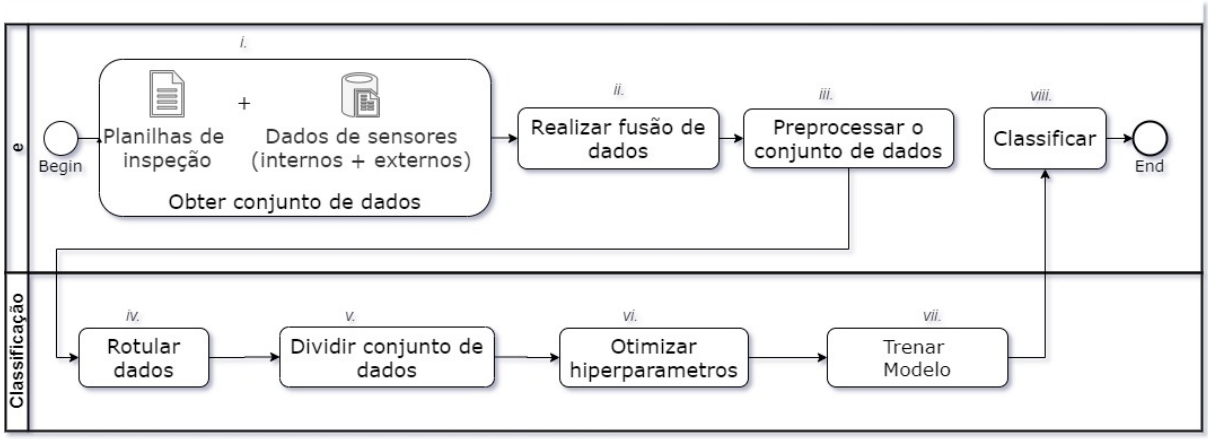
Por fim, vale destacar que nesse trabalho aplicou o agrupamento para encontrar níveis de bem estar das colônias, mas outros fenômenos relacionados às colônias de abelhas também poderiam ser investigados com o método aplicado para a busca de padrões de interesse, tais como abandono, enxameamento, perda da rainha e outros.

4 CLASSIFICAÇÃO

Este capítulo descreve os aspectos metodológicos e os resultados da abordagem que utiliza a técnica de mireação de dados classificação, detalhando tópicos relacionados às ferramentas utilizadas, coleta e pré-processamento de dados, bem como a análise e detecção dos níveis de bem estar de colônia de abelhas utilizando a classificação. Para a validação dessa abordagem, foi utilizado um conjunto de dados específico para essa abordagem conforme recomendado no Capítulo 1.

Nessa abordagem, os possíveis níveis de bem estar são definidos previamente, em termos quantitativos e qualitativos. As etapas a serem realizadas para obteção do modelo de classificação são ilustradas na Figura 23. A Seção 4.1 detalha cada uma dessas etapas. Lembrando que o processo ou método da Figura 23 é uma instância do método KDD apresentado no Capítulo2.

Figura 23 – Método de predição via aprendizado supervisionado (classificação)



Fonte: Elaborada pelo autor (2019)

4.1 Materiais e Métodos

Para validar do processo apresentado na seção anterior, foram usados dados da plataforma *Bayer Bee Care Center* (BBCC)[1]. O *Bee Care Center*[2] é uma plataforma criada para construir relacionamento com *stakeholders* com o objetivo de engajar e apoiar programas e estudos com universidades para a melhoria da saúde das abelhas.

[1] https://beecare.bayer.com/what-we-do/bayer-bee-care-program
[2] https://www.agro.bayer.com.br/beecare

4.1.1 Conjunto de dados

Foram utilizados dados de 27 colônias de abelhas (*Apis mellifera*) monitoradas entre janeiro de 2016 e dezembro de 2018. As informações meteorológicas das regiões das colmeias foram obtidas nas estações meteorológicas do Serviço Nacional de Meteorologia dos Estados Unidos, em inglês, *National Weather Service* (NWS). As colônias estão localizadas nos seguintes estados Norte Americanos: Indiana, Carolina do Norte, Pensilvânia e Utah.

4.1.1.1 Sensores internos

A leitura do sensor de temperatura na colmeia ocorria a cada 15 minutos pelo sistema de monitoramento de colmeias SolutionBee [3]. Este sistema é alimentado por bateria e essa taxa de amostragem foi definida com base nos requisitos de energia. Em cada colmeia, os dados foram baixados por Bluetooth para um celular e, em seguida, enviados para o armazenamento em nuvem fornecido pelo SolutionBee. O sensor de temperatura em todas as colmeias estava localizado nas barras superiores, perto do centro do ninho da colônia. A cada inspeção semanal, os sensores eram reposicionados para manter a proximidade do centro da área de criação. No total, foram coletadas 543.668 pontos amostrais em todas as colmeias. A Tabela 5 mostra os seis apiários usados, suas localizações, as coordenadas, o número de colmeias e o número de pontos amostrais por apiário.

Tabela 5 – Sumário dos apiários, quantidade de colmeias e pontos amostrais

Apiário	Localização	Lat.	Long.	#colmeias	#amostras
BBCC	Durham, NC	35.92°N	78.85°W	06	147799
BBTS	Pittsburg, PA	40.50°N	79.87°W	02	13870
Beesboro	Clayton, NC	35.64°N	78.43°W	08	92034
Juniper Level	Garner, NC	35.70°N	78.55°W	04	78751
Lakeview	Orem, UT	40.70°N	111.82°W	03	85068
The Bee Hive	Huntington, IN	40.80°N	85.55°W	04	126146

Fonte: Elaborada pelo autor (2019)

As seguintes variáveis foram monitoradas de cada colmeia:

1. Temperatura da colmeia em °Celsius: temperatura obtida com um sensor localizado dentro da colmeia;
2. Massa da colmeia Kg: peso da caixa + colônia obtido através de uma balança digital;

[3] https://solutionbee.com/

3. Variação da massa em Kg: um valor que indica se houve uma variação (positiva ou negativa) do peso da colmeia.

As Figuras 24(a) e 24(b) mostram, respectivamente, o apiário BBCC com quatro colmeias e o procedimento de sincronização de dados pelo apicultor *in loco*.

Figura 24 – Fotos do apiário BBCC sendo monitorado pelo SolutionBee

(a) colmeias com as balanças digitais

(b) Sincronização de dados

Fonte: Elaboradas pelo autor (2019)

4.1.1.2 Sensores externos às colmeias (dados meteorológicos)

A leitura dos sensores externo ocorria a cada 5 minutos. Contudo, para tornar o processo de treinamento dos algoritmos de classificação mais eficiente, uma média diária de cada sensor de dados externo foi calculada. A Tabela 6 mostra um resumo das cidades onde os conjuntos de dados meteorológicos foram obtidos com as respectivas fontes de dados, localização precisa da estação meteorológica (latitude e longitude) e a distância (em quilômetros) entre a estação meteorológica e o respectivo apiário.

Tabela 6 – Sumário das localizações das estações meteorológicas

Localização	Fonte de dados	Lat.	Long.	Distância
Durham, NC	Raleigh-Durham International Airport	35.89N	78.78W	7,89 km
Pittsburgh, PA	Pittsburgh International Airport	40.50N	80.27W	33,68 km
Clayton, NC	Johnston County Airport	35.54N	78.39W	12,42 km
Garner, NC	Johnston County Airport	35.54N	78.39W	23,58 km
Orem, Utah	Provo Municipal Airport	40.21N	111.71W	56,25 km
Huntington, IN	Fort Wayne International Airport	40.97N	85.21W	34,48 km

Fonte: Elaborada pelo autor (2019)

1. Temperatura externa (°Celsius): temperatura do ar;
2. Temperatura do ponto do orvalho (°Celsius): a temperatura em que o vapor de água no ar ambiente se torna líquido;
3. Direção do vento (graus): medido no sentido horário entre o norte verdadeiro e a direção na qual o vento está soprando;
4. Velocidade do vento (m/s): a taxa de deslocamento horizontal do ar além de um ponto fixo;
5. Precipitação em 1hr (milímetros): a espessura da precipitação líquida medida durante um período de acumulação de uma hora;
6. Luz do dia: uma variável categórica, indicando o nascer e o pôr do sol, calculada por meio de uma *Application Programming Interface* (API) que considera uma coordenada específica no globo e uma data.

4.1.1.3 Dados de inspeção

As inspeções *in loco* foram executadas uma por semana com o suporte HCC[4]. O Anexo A apresenta a referida lista de verificação. Essa lista verifica seis informações internas da colmeia a saber:

1. **Presença de todas as fases de cria** e ínstares em quantidades apropriadas;
2. **Presença de abelhas adultas suficientes** e estrutura etária para cuidar das crias e executar todas as tarefas da colônia;
3. **Presença de uma rainha jovem** ($\leq$ 1 ano de idade) e produtiva;
4. Alimentação suficiente: **água e pasto apícola** disponíveis (dentro e/ou fora da colmeia) e abelhas jovens sendo alimentadas;
5. **Presença de nenhum estressores (aparente)** que levariam a uma redução da sobrevivência da colônia e/ou no potencial de crescimento da colônia;
6. **Presença de espaço adequado** (não muito ou muito pouco) para o tamanho atual e de curto prazo esperado para o crescimento da colônia que garanta condições de higiene, de defesa e para a postura dos ovos.

Ao todo, 703 inspeções foram utilizadas neste trabalho. Uma validação dessas características foi feita por (JACOBS *et al.*, 2017). Para calcular a saúde da colônia, cada característica do HCC recebeu valores binários (0 ou 1). Se o item inspecionado não apresentava

[4] https://beehealth.bayer.us/who-can-help/beekeepers/healthy-colony-checklist

nenhum problema, o item recebia o valor 1, caso contrário, recebia o valor 0. Desses valores, a colônia foi categorizada em relação à sua saúde, usando uma soma desses valores para calcular o Estado de Saúde das colônias, ou do inglês, *Health Status* (HS) (Eq. 4.1). As inspeções foram feitas de maneira imparcial, uma vez que o HCC foi usado por apicultores experientes treinados para interpretar e preencher o HCC, que padronizou perguntas e itens a serem observados durante a inspeção. Além disso, cada condição foi considerada com base no que é considerado aceitável para cada estação do ano.

$$HS = \left(\frac{I1 + I2 + I3 + I4 + I5 + I6}{6} \right) \times 100 \tag{4.1}$$

onde $I1, I2, ..., I6$ são os valores de cada item de inspeção. Os níveis de saúde foram definidos pelo valor de HS. A partir de HS, foi possível definir seis níveis de estados de saúde das colônias (classes). Cada classe está associada a um nível de saúde. Portanto, se todos os itens estiverem marcados como "sem problemas"(valor 1), a colônia será considerada 100% íntegra ou com o estado de saúde 5. Por outro lado, cada item marcado com “problemas” (valor 0) representa uma redução de 1/6 no nível de saúde da colônia, ou seja, representa os estados de saúde 4, 3, 2, 1 ou 0.

4.1.2 Fusão dos dados de sensores

A fusão de dados de sensores internos e externos foi feita usando como referência a variável *timestamp* com uma janela de interseção de tempo de até 57 minutos, para mais ou para menos. Por exemplo, medições de sensores internos que ocorreram às 21:00:00 foram agrupadas com os dados de sensores externos que ocorreram entre 20:03:00 e 21:00:00. Assim, foi obtido um único arquivo com 485.371 pontos amostrais contendo as variáveis dos sensores interno e externo.

A seguir, usando a numeração apresentada no método da Figura 23, tem-se os seguintes passos: Rotular dados (iv); Dividir conjuntos de dados (v); Otimizar hiperparâmetros (vi).

4.1.3 Preprocessamento

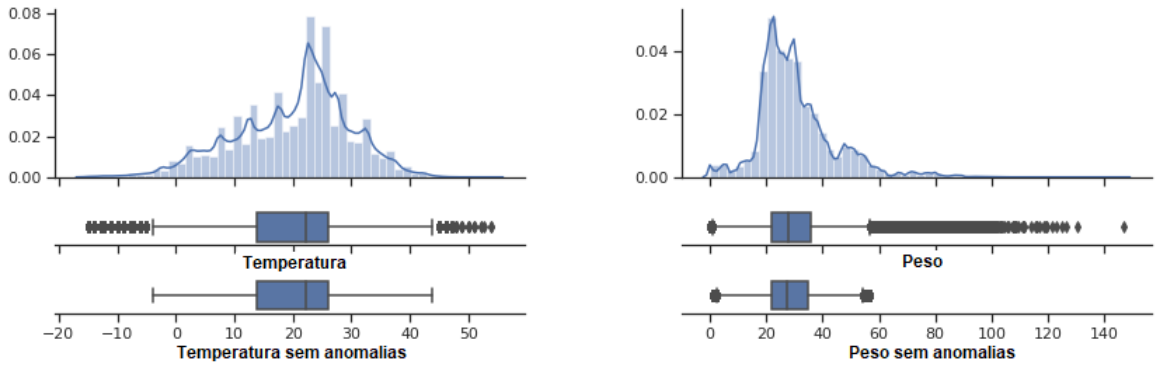
A segunda etapa da métodologia utilizada é o pré-processamento. O pré-processamento explora e analisa os dados para entender melhor o que eles representam. Muitas vezes, podem

ocorrer inconsistências, como dados discrepantes ou redundantes que dificultam o processamento e podem produzir um modelo preditivo impreciso. As principais etapas adotadas na fase de pré-processamento incluíram: fusão de dados detectados, análise exploratória de dados, detecção e remoção de outliers e redimensionamento de dados (padronização).

4.1.3.1 Análise exploratória de dados

O primeiro passo para familiarizar-se com os dados é conduzir uma análise inicial com as técnicas estatísticas básicas. Ou seja, a média, desvio padrão, intervalo interquartil e obliquidade. Além disso, a distribuição e o diagrama de caixa são realizadas para ver a distribuição dos valores de temperatura e peso. Esses gráficos podem ser observados na Figura 25(a), que mostra a Estimativa de Densidade por Kernel, em inglês, *Kernel Density Estimation* (KDE), e o gráfico de caixa de temperatura interna. A Figura 25(b) mostra o KDE e o gráfico de caixa do peso das colmeias.

Figura 25 – KDE e diagramas de caixas da temperatura e peso das colmeias



(a) Temperatura interna

(b) Peso das colmeias

Fonte: Elaboradas pelo autor (2019)

4.1.3.2 Detecção e remoção de valores discrepantes

Para o método baseado em classificação, a detecção de anomalias foi realizada usando dois métodos, um estatístico com o método Tukey (TUKEY, 1977) (conforme já apresentado na Seção 3.1.2.1) e um baseado em aprendizado de máquina com Fator Outlier Local, do inglês, *Local Outlier Factor* (LOF) (BREUNIG *et al.*, 2000).

O LOF é baseado no conceito de densidade local, em que a localidade é dada pelos

k vizinhos mais próximos, cuja distância é usada para estimar a densidade. Ao comparar a densidade local de um objeto com as densidades locais de seus vizinhos, pode-se identificar regiões de densidade semelhante e pontos que possuem uma densidade substancialmente menor que seus vizinhos. Estes últimos são considerados *outliers*. Assim,

- $LOF(k) \approx 1$ significa densidade Semelhante à dos vizinhos;
- $LOF(k) \leq 1$ significa densidade Maior que à dos vizinhos (*inlier*);
- $LOF(k) \geq 1$ significa densidade Menor que à dos vizinhos (*outlier*).

As Figuras 25(a) e 25(b) mostram os gráficos de caixa dos valores da temperatura interna e do peso da colmeia após a remoção dos *outliers*. Após a remoção dos valores discrepantes, o conjunto de dados ficou com 451.279 pontos amostrais.

4.1.3.3 Padronização de dados

A padronização do conjunto de dados foi realizada utilizando transformação *z-score* (Eq. 3.2), conforme já apresentado na Seção 3.1.2.2.

4.1.4 Rotulagem do conjunto de dados

Como os dados detectados não possuem rótulos, as classes foram obtidas na planilha de inspeção, através de um processo de sobreposição entre os dados de sensores e os dados de inspeção. Uma nova fusão foi feita através dos dados de inspeção. As classes foram atribuídas nos pontos amostrais de sensores imediatamente subsequentes a uma inspeção, de modo que permanecessem as mesmas entre uma inspeção e outra ou até que o estado da colônia mudasse. Essa atribuição resultou na distribuição observada na Tabela 7.

Como as inspeções ocorriam semanalmente, foi determinado um *threshold* de verificação da alteração do estado de saúde de 1 semana para mais ou para menos, assim, por exemplo, medições de sensores que ocorreram em $24/12/2019$ foram agrupadas com os dados de inspeção que ocorreram em uma semana a frente ou uma semana passada do dia da medição dos sensores.

Tabela 7 – Número de pontos amostrais por classe (estado de saúde).

Classe	0	1	2	3	4	5
Número de pontos amostras	442	564	1316	1276	8658	8562
Porcentagem de pontos amostras	2.12%	2.70%	6.32%	6.12%	41.58%	41.12%

Fonte: Elaborada pelo autor (2019)

Como pode ser visto, o número de pontos amostrais das classes ‘0’, ‘1’, ‘2’ e ‘3’ é pequena se comparada às classes ‘4’ e ‘5’. Esse problema é chamado de desequilíbrio de classe. O tratamento desse problema exigiu o uso de duas estratégias: (i) uso proporcional de pontos amostrais por classe nos conjuntos de treinamento, teste e validação (descritos em mais detalhes na seção 4.1.5) e (ii) o uso da métrica de avaliação *Area Under Receiver Operating Characteristic Curve* (AUC ROC) para avaliar a precisão dos modelos treinados (HE; GARCIA, 2009) (descrita em mais detalhes no Anexo B).

4.1.5 Divisão do conjunto de dados

Nesta fase, o conjunto de dados foi dividido nos subconjuntos de treinamento, validação e teste. A proporção escolhida foi 70:15:15. O subconjunto de treinamento foi usado para criar os modelos iniciais para cada algoritmo de classificação. O subconjunto de validação foi usado para ajustar os hiperparâmetros. A seção 4.1.6 descreve como esse ajuste foi executado para cada algoritmo. O subconjunto de testes é uma parte do conjunto de dados que é completamente nova para o modelo treinado. Os pontos amostrais do subconjunto de teste foram usadas para avaliar o desempenho dos classificadores.

Essa proporção também foi usada na distribuição dos pontos amostrais nos subconjuntos, ou seja, os pontos amostrais foram distribuídas aleatoriamente por subconjunto, proporcionalmente de acordo com o número de pontos amostrais por classe seguindo a proporção 70:15:15.

Além de dividir o conjunto de dados nesses três conjuntos, foi utilizada também a técnica *k-fold* de validação cruzada, em inglês *k-fold Cross-Validation*, para impedir que um única parte do conjunto de dados fosse usada. A CV foi configurado com 10 camadas, isto é, validação cruzada de 10 dobras.

4.1.6 Otimização dos hiperparâmetros

A otimização dos hiperparâmetros foi realizada de maneira específica para cada algoritmo na fase de validação. Para cada um, foi escolhida uma abordagem que melhor se ajusta à natureza do algoritmo.

Para o algoritmo kNN, o parâmetro de avaliação utilizado foi a precisão do método para diferentes valores de k e p, onde p é a ordem da distância de Minkowski. Em relação a p, foram avaliados os valores 1 e 2. A distância de Minkowski é calculada de acordo com a

Equação 4.2. Quando $p = 1$, a distância é equivalente à distância de Manhattan, quando $p = 2$, a distância é equivalente à distância euclidiana. Em relação a k, um vetor com valores para k foi definido no intervalo $[1,50]$. Todas as combinações possíveis de k e p foram testadas.

$$d_{XY} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{4.2}$$

Para o algoritmo RF, os hiperparâmetros avaliados foram o número de árvores ($trees$) e o número de atributos utilizados na criação de cada árvore ($mtry$). Para o $mtry$, testamos valores no intervalo $[1,9]$. Para o $trees$, de acordo com (OSHIRO *et al.*, 2012), uma boa aproximação para o número inicial de árvores para n atributos está entre 64 e 128. Para obter o valor de $trees$ no conjunto de validação, uma validação cruzada de 10 dobras foi executada.

Para o algoritmo NN, foi implementado uma rede MLP. A MLP é treinada interativamente, pois, a cada iteração, as derivadas parciais da função de perda são calculadas com relação aos parâmetros do modelo. O hiperparâmetro avaliado foi o termo de regularização L2 alfa (α). A função de ativação das camadas ocultas utilizadas foi a *Rectified Linear Unit* (ReLU), e a função de otimização de pesos foi a *Broyden-Fletcher-Goldfarb-Shanno Limited-memory* (L-BFGS). A arquitetura usada foi 1:5:2:1, ou seja, uma camada de entrada, duas camadas ocultas (com 5 e 2 neurônios) e uma camada de saída. Essa arquitetura foi obtida usando uma abordagem tentativa-erro, sendo que nos testes preliminar foi a arquitetura que apresentou melhores resultados. Contudo, orientações gerais para determinação da arquitetura de redes neurais foram obtidas em Hinton *et al.* (2006).

4.2 Resultados e Discussões

Para avaliar o desempenho do algoritmo de classificação, foram utilizadas seis métricas: (i) *Accuracy*, (ii) *Precision*, (iii) *Sensitivity* ou *Recall*, (iv) *F1-score*, (v) AUC ROC, and (vi) *Log loss*. Um detalhamento completo de cada métrica pode ser encontrado no Anexo B.

4.2.1 Configuração dos experimentos

A execução dos algoritmos de classificação foi dividida em quatro experimentos. Cada experimento foi realizado para definir a melhor configuração em relação ao número de classes e definir quais características são importantes para a definição do status de saúde

(estratégia conhecida por Seleção de Características ou *Feature Selection*). Cada experimento é descrito abaixo.

- Experimento #1: neste experimento, avaliou-se a precisão dos algoritmos, levando em consideração apenas duas condições de saúde da colônia: completamente saudáveis e com alguns problemas de saúde. Dessa maneira, só é possível identificar a transição do melhor estado de saúde para aquele com algum nível de doença. Esse tipo de classificação também é conhecido como classificação binária. Para realizar esta avaliação, as classes 0, 1, 2, 3 e 4 foram tratadas como uma só.
- Experimento #2: neste experimento, avaliou-se a precisão dos algoritmos, levando em consideração três status de saúde: saudável, fraco e doente. Para abordar essa avaliação, juntamos as classes 1, 2, 3 e 4. O objetivo deste experimento foi verificar se o uso de três classes era viável para um sistema de recomendação que busca prever um estado de doença iminente para que o apicultor tenha algum tempo para intervir na colônia de maneira que a mesma possa se recuperar.
- Experimento #3: neste experimento, avaliou-se a precisão dos algoritmos, levando em consideração quatro níveis de saúde: saudável, não saudável, fraco e doente. Para abordar essa avaliação, juntamos as classes 1 e 2. Da mesma forma, as classes 3 e 4 foram tratados como uma só. Assim, seria possível prever o estado de saúde de forma mais gradual. Com o uso de quatro classes, os modelos treinados poderiam ser usados em um sistema de recomendação capaz de sugerir ao apicultor iminentes problemas de saúde em uma colônia. Na prática, com essa divisão, o índice HS indicará uma colônia completamente saudável se todos os itens observados na inspeção não tiverem problemas. No entanto, se um ou dois itens apresentarem um problema, a colônia terá um problema de saúde. Quando três ou quatro itens apresentam um problema, a colônia está fraca. Finalmente, se cinco ou seis itens estiverem com problemas, a colônia estará em risco iminente de perda para o apicultor.
- Experimento #4: neste experimento, avaliou-se a precisão dos algoritmos levando em consideração todos os estados de saúde relatados pelo HCC, ou seja, seis estados de saúde. Assim, tentou-se avaliar a precisão dos algoritmos na predição de casos em que apenas um item do HCC apresentou problemas. Esse tipo de previsão tende a ser mais difícil para algoritmos, principalmente porque o conjunto de dados usado possui apenas o sensor de temperatura dentro da colmeia. No entanto, vale ressaltar que a métrica de temperatura é

crucial para o interior da colmeia (BECHER, 2010). Portanto, cada classe foi considerada individualmente na execução dos algoritmos.

Além de tentar definir adequadamente a quantidade de estados de saúde, foi executada também a *Seleção de Características*. O objetivo da seleção de características foi identificar a associação entre os atributos monitorados através de sensores e os itens observados pelo apicultor no HCC. Para isso, definiu-se quatro cenários, descritos abaixo, em que cada um possui uma combinação dos atributos obtidos pelos dados sensores, planilhas de inspeção e/ou outras medidas tomadas indiretamente (características extraídas). O objetivo de cada cenário é identificar como cada sensor e/ou atributo extraído contribuem para a definição do estado de saúde da colônia, a fim de responder perguntas, tais como: A temperatura interna da colmeia pode ser associada a algum item de inspeção da cria? A temperatura externa da colmeia pode ser associada a oferta de alimentos? O peso da colmeia pode ser associado ao espaço? Que sensor e/ou atributo extraído podem ser associados como o *status* da rainha ou dos estressores?

- Cenário #1: apenas sensores internos da colmeia foram utilizados no processo de classificação. Assim, os atributos (variáveis independentes) são a temperatura da colônia e o peso da colmeia.
- Cenário #2: foi configurado com dados internos e externos. Os dados internos são compostos de temperatura interna e peso da colmeia. Os dados externos utilizados foram: temperatura externa, ponto de orvalho, pressão, velocidade do vento e precipitação de 1 hora. Usando os dados externos, espera-se determinar a influência dos mesmos na saúde das colônias.
- Cenário #3: foi configurado com dados internos da colmeia, externos e informações sobre a rainha e os estressores obtidas através do HCC. Com essas informações, o objetivo foi avaliar se (i) os recursos do HCC também poderiam ser incluídos nos modelos de classificação e (ii) as informações sobre rainha e estressores poderiam melhorar a precisão do modelo, com base no pressuposto de que os outros atributos do HCC já estavam representados por dados de sensores e/ou por atributos extraídos.
- Cenário #4: foi configurado com informações internas, externas, rainha, estressoras e estações. Nesse cenário, o objetivo é saber como as estações do ano são relevantes na determinação da saúde das colônias.

Os resultados da execução dos algoritmos de classificação no conjunto de dados mostrado na Seção 4.1.1 podem ser vistos na Tabela 8. Para o experimento #1, os resultados mais

precisos foram obtidos nos cenários 3 e 4. Isso indica que os algoritmos foram capazes de prever com mais precisão os estados de saúde quando possuíam informações sobre temperatura interna, peso, rainha e estressores. No cenário 4, com a estação da ano, a taxa de acertos atingiu **98%** no algoritmo de RF. Além disso, a estação do ano não aumentou significativamente a precisão dos classificadores. No cenário 3, sem estação, a taxa de acerto era de 97% com o algoritmo RF. Esse resultado pode ser explicado pelo fato de a temperatura interna já ter uma forte correlação com a temperatura externa e, por sua vez, com a estação do ano. A taxa de acertos obtida nos cenários 1 e 2, foi menor que nos cenários 3 e 4, mostrando que, apenas com informações de sensores internos e externos, o algoritmo RF conseguiu prever o estado de saúde da colônia com uma taxa média de sucesso de até **91%**, que pode ser considerada uma alta taxa de acerto.

Com relação aos algoritmos executados, é possível observar convergência nos resultados, indicando que os ajustes realizados nos hiperparâmetros foram realmente os melhores possíveis. Também é possível observar que a precisão dos três algoritmos é semelhante, mesmo que eles sejam baseados em três algoritmos essencialmente diferentes. A precisão dos algoritmos é corroborada por todas as outras métricas calculadas. No entanto, em todos os cenários, o algoritmo de RF apresentou melhores taxas de certeza, com ênfase na métrica *Log Loss* do RF que indica alta confiabilidade do modelo.

Para o experimento #2, como pode ser visto na Tabela 8, novamente, os cenários 3 e 4 mostraram os melhores resultados com uma taxa de acerto **98%** com o algoritmo RF. Isso reforça a hipótese de que os itens do HCC relativos à rainha e estressores não estão sendo monitorados diretamente. Se comparado ao experimento #1, em relação ao uso de três classes, é possível observar que a taxa de acerto não mudou. Nos cenários 1 e 2 (apenas sensores), a taxa média de acerto foi reduzida em 8%, com o melhor resultado no cenário 2, com o algoritmo de RF, uma taxa de acerto de **90%**. Em suma, com a prerrogativa de usar um nível intermediário de estado de saúde para alertar o apicultor a tempo de fazer algum manejo preventivo na colônia, é possível usar três classes com uma precisão média de 90%. Novamente, isso pode ser considerado uma alta taxa de acerto. Em relação aos algoritmos testados, o RF novamente apresentou melhores taxas de acerto quando comparado ao kNN e ao NN, no entanto, com taxas próximas. Todas as outras métricas calculadas também reforçaram a precisão calculada.

Para o experimento #3, como pode ser visto na Tabela 8, os cenários 1 e 2 tiveram uma taxa de acerto menor quando comparados aos experimentos #1 e #2. Assim, os algoritmos tiveram mais dificuldade em prever o estado de saúde com o uso de dois níveis intermediários

Tabela 8 – Resultados dos experimentos

Experimento	Cenário	Algoritmo	Métricas					
			Accuracy	Precision	Recall	F1-score	AUC	Log Loss
1	1	kNN	0.84	0.85	0.83	0.84	0.83	0.63
		RF	0.88	0.88	0.88	0.88	0.88	0.37
		NN	0.84	0.84	0.84	0.84	0.84	0.35
	2	kNN	0.83	0.83	0.82	0.83	0.82	0.67
		RF	**0.91**	**0.92**	**0.91**	**0.91**	**0.91**	**0.24**
		NN	0.85	0.84	0.84	0.85	0.84	0.34
	3	kNN	0.93	0.94	0.93	0.93	0.93	2.27
		RF	0.97	0.97	0.98	0.97	0.97	0.09
		NN	0.95	0.95	0.95	0.95	0.95	0.12
	4	kNN	0.95	0.95	0.96	0.95	0.95	1.66
		RF	**0.98**	**0.98**	**0.99**	**0.98**	**0.98**	**0.07**
		NN	0.96	0.96	0.96	0.96	0.96	0.08
2	1	kNN	0.84	0.84	0.81	0.84	0.85	5.59
		RF	0.89	0.89	0.88	0.89	0.90	0.35
		NN	0.84	0.85	0.76	0.84	0.83	0.38
	2	kNN	0.82	0.83	0.71	0.81	0.79	0.69
		RF	**0.90**	**0.93**	**0.87**	**0.90**	**0.90**	**0.27**
		NN	0.84	0.85	0.78	0.84	0.84	0.35
	3	kNN	0.93	0.92	0.91	0.93	0.93	2.40
		RF	0.97	0.97	0.96	0.97	0.97	0.10
		NN	0.95	0.94	0.93	0.95	0.95	0.12
	4	kNN	0.95	0.96	0.93	0.95	0.95	1.67
		RF	**0.98**	**0.98**	**0.96**	**0.98**	**0.98**	**0.08**
		NN	0.96	0.96	0.94	0.96	0.96	0.09
3	1	kNN	0.79	0.76	0.76	0.79	0.84	7.22
		RF	0.85	0.84	0.83	0.85	0.88	0.50
		NN	0.77	0.74	0.67	0.77	0.79	0.52
	2	kNN	0.79	0.75	0.74	0.79	0.83	7.40
		RF	**0.88**	**0.90**	**0.84**	**0.88**	**0.89**	**0.33**
		NN	0.79	0.79	0.71	0.79	0.81	0.48
	3	kNN	0.92	0.91	0.90	0.92	0.93	2.90
		RF	0.96	0.96	0.96	0.96	0.97	0.12
		NN	0.93	0.93	0.92	0.93	0.95	0.16
	4	kNN	0.93	0.92	0.92	0.93	0.94	2.33
		RF	**0.97**	**0.96**	**0.95**	**0.97**	**0.97**	**0.11**
		NN	0.95	0.94	0.93	0.95	0.95	0.11
4	1	kNN	0.77	0.69	0.70	0.77	0.81	7.81
		RF	0.82	0.76	0.74	0.82	0.85	0.60
		NN	0.74	0.68	0.58	0.73	0.75	0.62
	2	kNN	0.76	0.70	0.68	0.76	0.81	8.28
		RF	**0.87**	**0.85**	**0.80**	**0.87**	**0.88**	**0.40**
		NN	0.75	0.73	0.60	0.74	0.77	0.58
	3	kNN	0.89	0.84	0.84	0.89	0.90	3.78
		RF	0.95	0.94	0.90	0.94	0.95	0.19
		NN	0.91	0.87	0.84	0.91	0.91	0.21
	4	kNN	0.91	0.84	0.85	0.91	0.91	0.31
		RF	**0.97**	**0.95**	**0.94**	**0.97**	**0.97**	**0.13**
		NN	0.93	0.89	0.89	0.93	0.94	0.16

Fonte: Elaborada pelo autor (2019)

de saúde, ou seja, com maiores detalhes da evolução da saúde da colônia. No experimento #3, a diferença de precisão entre os cenários 1 e 2 e os cenários 3 e 4 foi maior, evidência de que os sensores de temperatura e peso não podem monitorar diretamente a atividade da rainha e a presença de estressores. No entanto, a taxa de acertos nos cenários 1 e 2 ainda pode ser considerada muito boa (**85%**). No cenário 4, a taxa de acertos permanece alta (**97%**). Portanto,

o uso de duas classes intermediárias para as classes extremas também é uma configuração viável para um sistema de recomendação para o apicultor.

Com relação aos algoritmos testados neste experimento (com 4 classes), o RF também produziu os resultados mais precisos para a métrica "*Accuracy*"(precisão geral) e outras métricas baseadas em positivo verdadeiro, falso positivo, verdadeiro negativo e falso negativo (*Precision, Recall, F1- score e AUC ROC*). O RF também foi o mais preciso em relação à confiabilidade das previsões feitas (*Log Loss*), pois apresentou o menor valor quando comparado aos demais algoritmos.

Para o experimento #4, como pode ser visto na Tabela 8, a taxa de acertos nos cenários 1 e 2 foi menor quando comparada aos experimentos #1, #2 e #3. No cenário 2, a taxa de acerto foi **87%**, o que demonstra a capacidade do RF de distinguir e prever o status de saúde da colônia usando seis classes com os sensores usados. No cenário 4, a taxa de acerto de maior precisão foi **97%**, resultado muito semelhante ao observado em experimentos anteriores. Assim, o uso de seis classes de saúde também é viável.

Como nos outros experimentos, a RF foi o algoritmo que apresentou a melhor taxa de acerto, considerando todas as métricas e cenários. No caso do cenário 4, o *Log Loss*, por exemplo, apresentou um valor baixo, o que indica que o modelo criado pelo RF era confiável. Essa confiabilidade do modelo foi reforçada pelas outras métricas calculadas. Assim, é possível concluir que o algoritmo de RF é o mais apropriado para a classificação dos estados de saúde das colônias de abelhas. Na Tabela 9, é possível observar os valores dos hiperparâmetros calculados pela etapa de otimização para cada experimento/cenário/algoritmo.

Tabela 9 – Configuração dos hyperparâmetros

Algoritmo	Hyp	Experimento #1				Experimento #2				Experimento #3				Experimento #4			
		Cenários															
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
kNN	k	8	9	1	1	1	11	1	1	1	1	1	1	1	1	1	1
	p	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
RF	trees	75	92	98	84	86	86	94	84	86	84	77	88	88	96	96	83
	mtry	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
NN	α	1^{-4}	1^{-3}	1^{-5}	1^{-5}	1^{-2}	1^{-3}	1^{-4}	1^{-5}	1^{-3}	1^{-2}	1^{-6}	1^{-3}	1^{-4}	1^{-2}	1^{-3}	1^{-4}

Fonte: Elaborada pelo autor (2019)

Nos quatro experimentos, o uso de apenas sensores internos e externos apresentou uma taxa de sucesso muito boa, 91%, no experimento #1. Este é um indicador relativamente confiável da saúde da colônia, embora com apenas dois estados de saúde (totalmente saudáveis e

com algum problema de saúde). O uso de um estado intermediário de saúde (experimento #2) também se mostrou um cenário confiável a 90%, no cenário 2. Nos cenários 3 e 4, também foi possível observar que nenhum dos sensores usados monitorava diretamente o atividade da rainha e presença de estressores.

Em relação ao atributo estação do ano, embora alguns estados de saúde sejam recorrentes em algumas estações (por exemplo, colônias doentes no inverno), a indicação da estação para os algoritmos de classificação não foi decisiva no cenário 4. Em outras palavras, a capacidade das colônias de manter a termorregulação (temperatura interna), o número de reservas alimentares e o tamanho da colônia (peso da colmeia) foram mais importantes na determinação do estado de saúde, bem como da atividade da rainha e da presença de estressores, como já mencionado (BRAGA *et al.*, 2019).

A importância da inspeção padronizada reflete-se no processo de treinamento dos algoritmos de classificação, que atingiram uma alta taxa de sucesso, mesmo com as inspeções realizadas por diferentes apicultores nas colmeias localizadas em regiões distantes e diferentes. Assim, com a padronização, o RF, por exemplo, conseguiu criar árvores que representam os mais diversos casos associados ao estado de saúde das colônias.

A Tabela 10 mostra a proporção de pontos amostrais ao longo dos anos por estado de saúde. Como pode ser visto, o percentual de colônias completamente saudáveis (classe 5), ou com apenas um item de inspeção com problema (classe 4), aumentou ao longo dos anos. Esse percentual foi de 76,66% em 2016, foi de 75,41% em 2017 e atingiu 91,27% em 2018. Comparando com o teste de diferença de proporções, com nível de confiança de 99%, 2016 versus 2017 não apresentou diferenças significativas (p-*value*$= 0,1605$). Por outro lado, comparando 2016 versus 2018, houve diferenças significativas (p-*value* $< 0,001$) e 2017 versus 2018 o mesmo (p-*value* $< 0,001$), ou seja, há evidências de que houve uma melhora significativa em 2018. Essa melhoria está associada a inspeções padronizadas ao longo dos anos, que contribuem para manter a saúde das colônias e se torna significativamente importante se considerarmos o aumento no número de colônias inspecionadas/monitoradas, de 8 em 2016 no início do programa de inspeção, a 18 colônias em 2018. Em 2017, 12 colônias foram inspecionadas (veja Tabela 10).

A distribuição dos pontos amostrais ao longo dos meses dos anos pode ser vista na Figura 26. O período de inverno, de dezembro a março, é um período importante a ser observado neste gráfico, porque é especialmente difícil realizar inspeções. O ano de 2016 não possui pontos amostrais de inverno; no entanto, é possível observar que os estados menos saudáveis ocorrem

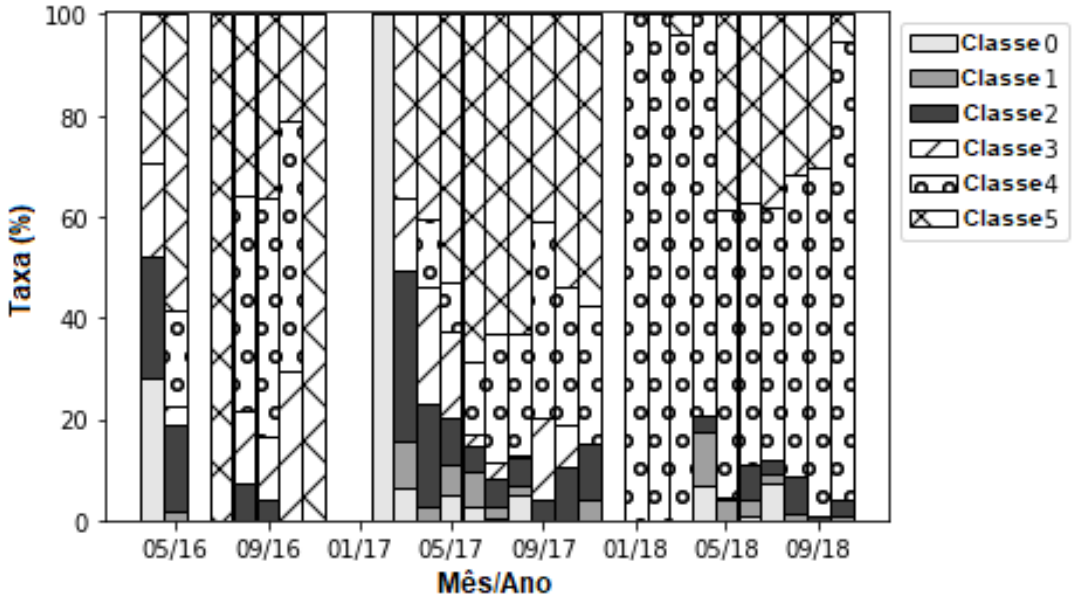
Tabela 10 – A proporção de estados de saúde ao longo dos anos observados

Ano	Classe						#colonias
	0	1	2	3	4	5	
2016	1.23%	0.18%	5.38%	16.52%	40.09%	36.57%	8
2017	2.66%	3.53%	9.40%	8.97%	20.17%	55.24%	12
2018	1.95%	2.85%	3.90%	0.00%	61.10%	30.17%	18

Fonte: Elaborada pelo autor (2019)

principalmente nos meses de abril e maio e que as colônias se tornaram mais saudáveis nos meses seguintes.

Figura 26 – Proporção dos estados de saúde por mês



Fonte: Elaborada pelo autor (2019)

Em fevereiro, março, abril e maio de 2017, é possível observar uma maior concentração de casos de colônias doentes. No entanto, após o inverno e com o início do período de forrageamento na primavera, pode-se observar uma diminuição nos casos de colônias doentes. O benefício do programa de inspeção é ainda mais evidente durante e após o inverno de 2017-2018. Em janeiro e fevereiro de 2018, todas as colônias estavam em estado de saúde 4, o que é muito bom. Uma baixa incidência de colônias doentes pode ser observada em abril de 2018. No entanto, esse número foi reduzido nos meses seguintes. Em Stalidzans e Berzonis (2013), é possível observar um estudo em que os autores mostram que os períodos de desenvolvimento de colônias podem ser previstos usando sensores de temperatura, de acordo com um ciclo anual.

Também é importante observar quais itens do HCC são mais relatados como "pro-

blemáticos"pelo apicultor. A Tabela 11 mostra a porcentagem de ocorrência de itens de HCC identificados pelo apicultor como tendo alguns problemas, ou seja, na planilha de inspeção, os problemas foram marcados com '0'. Como se pode observar, os itens 'Rainha' e 'Estressor' foram os itens que mais pareceram problemáticos nas inspeções, com 27,02% e 31,54%, respectivamente. A recorrência desses itens sugere que eles devam receber atenção especial no gerenciamento e no uso das melhores práticas para evitar problemas. No caso da administração da "rainha", por exemplo, os apicultores devem inspecionar regularmente (a frequência depende da estação), limitar a inspeção durante o inverno e com mau tempo, marcar a rainha com uma cor específica para o ano de sua introdução no colmeia e substituição da rainha a cada 2 ou 3 anos (de preferência no final do verão). No caso dos estressores, por exemplo, os apicultores devem sincronizar as medidas de controle da Varroa e, de preferência, setorizando a aplicação de remédios por regiões para um melhor controle (FORMATO; SMULDERS, 2011).

Tabela 11 – Proporção de ocorrência de itens de inspeção

	Cria	Abelhas	Rainha	Alimentos	Estressores	Espaço
Quantidade	3050	1546	5626	1956	6568	2698
Percentual	14.65%	7.43%	27.02%	9.39%	31.54%	12.95%

Fonte: Elaborada pelo autor (2019)

Os itens 'Cria' e 'Espaço' apareceram de maneira moderada, em 14,65 % e 12,95 %, respectivamente. O item 'Cria' pode sofrer indiretamente com a ocorrência do problema nos itens 'Rainha' e 'Estressores'. Os itens 'Abelhas' e 'Alimento' tiveram o menor número de incidentes, com 7,43 % e 9,39 %, respectivamente. Esse achado sugere que, quando as abelhas atingem a fase adulta, elas tendem a sofrer menos problemas de saúde.

Vale ressaltar que essas proporções foram calculadas considerando todos os apiários durante todo o intervalo de tempo observado nos experimentos. No entanto, uma análise semelhante pode ser feita de maneira mais segmentada, pois essas proporções podem variar entre apiários e época do ano. Outro ponto importante diz respeito ao processo de treinamento dos algoritmos: mesmo com alta incidência dos itens 'Rainha' e 'Estressor' e sem detecção direta desses itens, o algoritmo de RF conseguiu identificar as altas taxas de acerto dos estados de saúde.

A partir das altas taxas de acerto obtidas nos algoritmos de classificação, foi possível também buscar informações práticas no conjunto de dados sobre as temperaturas externa e interna das colmeias. De acordo com Heldmaier (1987), a capacidade da colônia de sobreviver

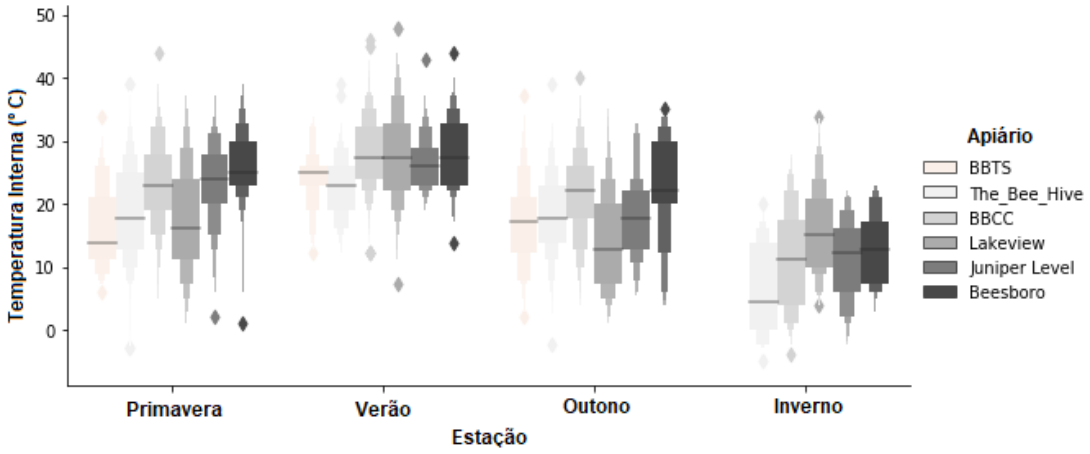
ao frio depende da manutenção de uma temperatura de estado normal, cerca de 35°C, na área central da colônia. No entanto, dado que em algumas regiões a temperatura externa pode atingir menos que 0°C no inverno, isso pode ser um complicador que podem levar a vários problemas na saúde da colônia.

Caso a temperatura interna seja muito baixa, a conseqüência é que a cria morre e o néctar não desidrata rápido o suficiente para produzir mel. Em clima extremamente frio, as abelhas adotam um comportamento protetor e iniciam um fenômeno chamado diapausa. É uma interrupção gradual e progressiva no desenvolvimento ou na ongenia das abelhas, para que elas possam sobreviver a essas condições ambientais desfavoráveis (WINSTON, 1991). Altas temperaturas são outro problema para a saúde das abelhas. O superaquecimento ocorre quando a temperatura interna da colmeia excede 36 °C, o que pode levar à morte da cria, com a fusão da cera e a desidratação do mel muito rapidamente (WINSTON, 1991).

No caso dos apiários estudados neste trabalho, eles estão localizados em cidades diferentes e as temperaturas internas de cada apiário têm padrões diferentes entre as estações do ano. Como pode ser visto na Figura 27, na primavera, a temperatura média variou entre 15° e 25° entre os apiários, os apiários 'Beesboro', 'Juniper Level' e 'BBCC' com medianas semelhantes, cerca de 24°C, porque eles estão no mesmo estado, Carolina do Norte. Nos apiários 'BBTS', 'The_Bee_Hive' e 'Lakeview' a temperatura interna média nas colmeias era de cerca de 15°C. No verão, a variabilidade entre os apiários foi a mais baixa, com a menor mediana de 22°C no apiário 'The_Bee_Hive'. Todos os outros apiários tinham uma temperatura média interna em torno de 25°C. No outono, os apiários localizados na Pensilvânia e Indiana, 'BBTS' e 'The_Bee_Hive', respectivamente, tiveram valores de temperatura interna semelhantes, cerca de 15°C. Os outros apiários (BBCC, Lakeview, Juniper Level e Beesboro) variaram com temperaturas entre 10° e 21°C. No inverno, a temperatura média não era superior a 13°C, com a menor mediana de 0°C no apiário 'The_Bee_Hive'. No apiário 'BBCC', a temperatura média era de cerca de 0°C e nos apiários 'Beesboro', 'Juniper Level' e 'Lakeview' era de cerca de 13°C. O apiário 'BBTS' não teve ponto amostral durante o inverno. Como esperado, as temperaturas internas estão fortemente associadas à temperatura externa, conforme discutido em mais detalhes por Rice (2013). É importante notar que, mesmo no inverno em que a temperatura interna é mais baixa, as colônias podem manter um bom estado de saúde se mostrarem todos os itens de inspeção sem problemas (ABOU-SHAARA *et al.*, 2017).

Em relação à temperatura interna e ao estado de saúde das colmeias (Figura 28), em

Figura 27 – Temperatura interna das colmeias por estação e apiário

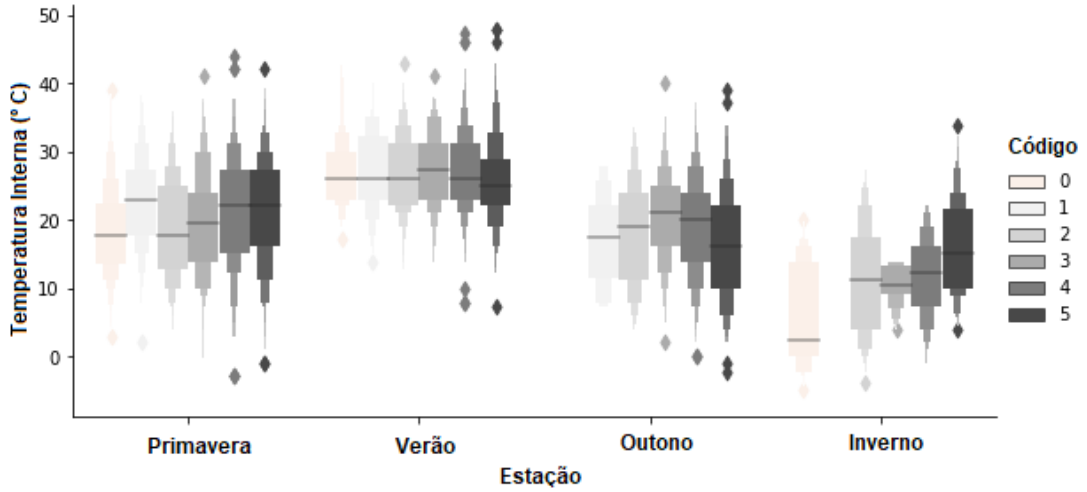


Fonte: Elaborada pelo autor (2019)

todas as estações do ano, no conjunto de dados analisado, foi possível extrair uma tendência de temperaturas mais altas nas classes mais saudáveis. Na primavera, as classes que representavam os piores estados de saúde (0, 1 e 2) tinham uma temperatura interna em torno de 18°C e as classes dos estados mais saudáveis (4 e 5) em torno de 22°C. No verão, não foi possível observar uma grande diferença de temperatura entre os estados de saúde (cerca de 25°C), principalmente porque a temperatura interna é geralmente mais alta no verão (ABOU-SHAARA *et al.*, 2017), pois o controle térmico depende da presença de um número maior de abelhas trabalhadoras saudáveis e reservas de mel (COOK; BREED, 2013; SÁNCHEZ *et al.*, 2015). No outono, foi possível extrair um padrão semelhante de distribuição de temperatura ao observado na primavera. Como as temperaturas externas começaram a cair no outono, a temperatura interna da colmeia dependia do estado de saúde da colônia, porque colônias fracas não são capazes de termorregular. Assim, novamente, os estados menos saudáveis têm temperaturas internas mais baixas, em torno de 17° e 19 °C. Finalmente, no inverno, foi possível observar que as colônias são muito suscetíveis à variação de temperatura. Nos estados menos saudáveis, a temperatura interna está entre 3° e 10°C, e nos estados mais saudáveis, em torno de 10° e 15°C. Meikle *et al.* (2017) observaram que a falta de controle da temperatura interna está diretamente associada à saúde da colônia. Os autores observaram ainda que, antes e durante o inverno, as colmeias utilizadas na polinização comercial eram doentes com um controle de temperatura reduzido.

Adicionalmente, o escore do estado de saúde pode ser considerado tendencioso, pois seu cálculo não é ponderado por fatores como condição da rainha, estado da cria e estressores. Dado que não há base empírica consistente para a ponderação dos fatores no momento, no presente estudo optou-se por evitar influências artificiais nos resultados e por testar diretamente

Figura 28 – Temperatura interna por estado de saúde



Fonte: Elaborada pelo autor (2019)

a proeminência dos fatores intuitivos. Isso se reflete nos cenários 3 e 4 do projeto experimental.

Adicionalmente, a fim de exlorar como a composição dos fatores poderia ser feitas no sentido de criar uma base empirica para ponderação dos fatores, experimentos adicionais foram realizados com 4 colmeias do conjunto de dados apresentado na Seção 4.1.1 (BRAGA *et al.*, 2019), são elas: duas do apiário 'BBCC', em Durham, monitoradas no periodo de Junho de 2017 à Abril de 2019, e 2 no apiário Beesboro em Clayton, monitoradas no periodo de Junho de 2017 à Junho de 2018. Nessa abordagem, o agrupamento foi usada para determinar uma quantidade otimizada de níveis de saúde e a classificação para criação de um modelo de predição. Na etapa de agrupamento, o índice de validação CH e o algoritmo *k-means* foram usados para determinar a quantidade ideal de classes de saúde das colônias. Na classificação, foram treinados 4 algoritmos de classificação distintos, o kNN, o RF, as NN e o SVM.

Para melhor ilustrar a abordagem desse experimento, a seguir, serão apresentados dois algoritmos que mostram as etapas executadas. O Algoritmo 2 nos fornece o melhor valor do número de grupos com base no índice CH, em síntese, esse algoritmo recebe vários candidatos ao número de grupos e para cada candidato o laço for (L. 2-9) agrupa o conjunto de dados de interesse pelo algoritmo *k-means* (L. 3-7). Para esse candidato em específico é calculado o índice CH associado (L. 8). Então, o melhor valor do número de grupos é aquele com maior valor associado ao vetor de índices CH.

O Algoritmo 3 utiliza o resultado do Algoritmo 2 para realizar uma busca exaustiva

Algoritmo 2: Escolha do melhor valor de k via *k-means* e o índice CH

Data: **K** (vetor com os possíveis números de grupos), **D** (um conjunto de dados)
Result: Um vetor com os índices CH associados ao vetor **K**

```
1  indices = {} ;
2  foreach k ∈ K do
3      escolha k observações de D como os centróides iniciais;
4      repeat
5          atribua cada observação de D ao grupo com maior similaridade, baseada na
             média dos objetos no grupo;
6          atualize as médias em cada grupo com as observações realocadas;
7      until convergência;
8      insira em indices o índice de CH associado aos k grupos;
9  end
10 retorne indices;
```

do melhor agrupamento, em síntese, esse algoritmo recebe um valor ótimo k do número de grupos, fatores de inspeção e um conjunto de dados. É inserido, então, todos os possíveis agrupamentos em k grupos dos fatores de inspeção (L. 2). No laço for (L. 4-9) é realizada a rotulagem do conjunto de dados e o cálculo da acurácia de um classificador genérico. A rotulagem (L. 6) é realizada através de uma associação direta entre a quantidade de itens de inspeção saúdáveis conforme já apresentado na Seção 4.1.4. Ao final do algoritmo, é retornada um vetor com o desempenho/acurácia de cada agrupamento de k grupos pelo modelo de classificação escolhido. A partir desses resultados pode se escolher o melhor agrupamento de fatores e o melhor algoritmo a ser utilizado.

Em cada avaliação do classificador $j \in \{\text{kNN, RF, NN e SVM}\}$, é feito um experimento de validação cruzada de 10-dobras (10-fold) e avaliado a acurácia com um vetor H_j de hiperparâmetros possiveis. O vetor H_j para cada classificador é definido na Tabela 12,

Para execução do Algoritmo 2, foi utilizado um vetor K para os possíveis números de grupos onde $K = 3, ..., 10$. Essa definição se deu com o objetivo de caracterizar no mínimo à partir de 3 estados de saúde, que podem ser entendidos como: saudável, alerta e doente. Para $k = 2$ obtem-se uma classificação binária, em que as duas possíveis classes são extremas e, portanto, não há uma classe que sirva de alerta para o usuário final. O resultado pode ser observado na Figura 29.

O melhor valor de k é 3. Após a definição do número de classes ($k = 3$), foi possível executar o Algoritmo 3 a fim de definir o melhor agrupamento de fatores de inspeção. O número

Algoritmo 3: Escolhe melhor agrupamento dos fatores de inspeção

Data: **k** (quantidade de classes desejadas), **F** (fatores de inspeção), **D** (conjunto de dados correspondente à inspeção), **C** (classificadores)

Result: Um vetor de acurácias obtidas por um classificador genérico associadas a cada possível agrupamento de tamanho **k**

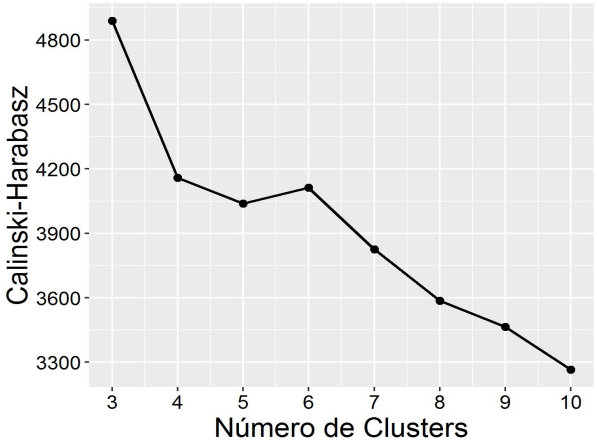
1 **acurácias** = {};
2 **agrupamentos** = possíveis agrupamentos de fatores;
3 **soma** = soma dos fatores de inspeção;
4 **foreach** *agrupamento* ∈ *agrupamentos* **do**
5 **classe** = { } ; `/* classe ou estado de saúde a ser atribuído */`
6 **atribua** o valor da classe (0 a **k**) à variável **classe**, baseada na soma dos fatores de inspeção **F** e no agrupamento da iteração atual;
7 **treine** o classificador genérico **C** utilizando o conjunto de dados **D** combinado à **classe**, como preditora, em um experimento de validação cruzada;
8 **insira** em **acurácias** a métrica acurácia correspondente ao experimento de validação cruzada na iteração atual;
9 **end**
10 **retorne acuracias**;

Tabela 12 – Configurações dos classificadores: k-NN, RF, NN e SVM

Classificador	Hiperparâmetros Fixos	Hiperparâmetros possíveis
kNN	Nenhum	k: {1,2, ..., 30}
RF	Nenhum	mtry: {2, 7, 12}, splitrule: {gini, extratrees}
NN	MLP, dim: (400, 200), dropout: 0.45, batch_size: 100, optimizer: Adam	Nenhum
SVM	kernel: RBF, gamma: 1.5, C = 9	Nenhum

Fonte: Elaborada pelo autor (2019)

Figura 29 – Gráfico resultante da aplicação do Algoritmo 2



Fonte: Elaborada pelo autor (2019)

de possíveis agrupamentos é igual 90. Os resultados são mostrados na Tabela 13, onde a coluna "*Agrupamento*" fornece a identificação de cada possível agrupamento.

Os agrupamentos estão ordenados de acordo com os maiores mínimos de sensibilidade e especificidade interclasses. Essa abordagem nos proporcionará uma configuração de grupos que seja menos penalizada pelos algoritmos no experimento e com alta acurácia. Para escolher o melhor agrupamento tomaremos como base um *trade-off* entre acurácia, sensibilidade e especificidade. Embora, o agrupamento 2456 1 3 nos forneça a maior acurácia dos agrupamentos (99.36%), esse resultado não é o melhor, pois, o que ocasiona essa alta acurácia é o excesso de fatores agrupados. O mesmo ocorre para os agrupamentos 1456 2 3 e 3456 1 2.

Tabela 13 – Acurária dos agrupamentos para os classificadores: kNN, RF, NN e SVM

	Acurácia (%)			
Agrupamento	kNN	RF	NN	SVM
5 12 346	90.11	95.09	91.48	90.73
2 13 456	**98.55**	**99.21**	**99.23**	**98.32**
12 35 46	90.27	94.89	92.55	90.67
2 15 346	89.87	94.89	90.90	90.59
13 25 46	90.18	94.80	91.84	90.21
15 23 46	89.75	94.83	91.61	90.50
1246 3 5	90.50	95.04	91.22	91.12
2 46 135	89.88	94.83	91.61	90.31
13 24 56	94.29	96.69	95.35	94.78
5 23 146	89.57	94.80	91.44	90.21
5 46 123	90.19	95.15	91.97	90.61
2 35 146	89.77	94.88	90.27	90.37
5 13 246	89.95	95.02	92.17	90.85
1456 2 3	**98.50**	**99.11**	**98.60**	**98.49**
2 56 134	94.04	96.65	95.85	94.75
3 56 124	94.56	96.79	96.08	95.07
3 15 246	89.88	94.87	91.54	90.47
1346 2 5	89.97	94.93	92.27	90.71
3 25 146	89.72	94.67	92.40	90.21
2 34 156	94.06	96.65	95.09	94.63
3 24 156	94.26	96.73	95.22	94.76
3 12 456	**98.75**	**99.25**	**99.16**	**98.61**
12 34 56	94.29	96.76	96.30	94.69
3 46 125	90.20	94.87	91.09	90.42
5 34 126	88.88	94.23	90.73	90.01
3 45 126	92.07	95.42	92.96	92.57
1 23 456	**98.52**	**99.12**	**99.16**	**98.33**
14 23 56	93.94	96.56	95.72	94.28
3 14 256	93.95	96.45	95.44	94.54
2 14 356	93.93	96.62	95.50	94.51
4 35 126	88.99	94.13	90.96	89.83
4 23 156	93.91	96.71	95.57	94.25
4 12 356	94.34	96.73	95.65	94.62
2456 1 3	**99.06**	**99.36**	**99.16**	**98.99**
4 13 256	94.00	96.64	95.39	94.38
4 56 123	94.26	96.80	95.80	94.59
1256 3 4	94.28	96.66	95.76	94.86
5 14 236	88.53	94.00	90.40	89.55
4 15 236	88.40	94.09	89.93	89.59
1236 4 5	89.13	94.19	90.85	89.94
4 25 136	88.63	93.98	90.73	89.37
14 26 35	88.66	93.99	90.98	89.50
1 24 356	94.31	96.83	95.44	94.84
2356 1 4	94.15	96.58	95.59	94.60
5 26 134	88.82	94.17	91.18	90.21
1356 2 4	93.82	96.53	95.93	94.49
3 26 145	91.83	95.29	93.80	92.38
5 24 136	88.78	94.22	91.03	89.95
2 45 136	91.89	95.40	93.44	92.26
13 26 45	91.80	95.28	93.37	92.36
3456 1 2	**98.68**	**99.32**	**99.03**	**98.72**
1 34 256	94.04	96.58	95.01	94.74
1 56 234	94.42	96.88	95.93	95.16
5 16 234	89.05	94.25	91.33	90.08
4 26 135	88.57	94.00	91.39	89.59
2 36 145	91.87	95.27	93.01	92.50
3 16 245	92.29	95.37	93.89	92.89
5 36 124	89.27	94.17	92.43	90.37
15 26 34	88.53	94.15	90.83	89.70
1 35 246	89.95	94.98	91.80	90.82
1 45 236	91.92	95.28	93.37	92.14
1 26 345	92.01	95.43	93.72	92.67
1 36 245	92.27	95.48	93.89	92.71
1 46 235	90.15	94.62	91.22	90.76
6 25 134	88.97	94.24	90.85	89.81
1235 4 6	89.18	94.06	90.75	90.04
1234 5 6	89.62	94.42	91.67	90.62
6 12 345	92.52	95.54	93.69	92.81
15 24 36	88.77	94.30	90.42	89.83
6 35 124	89.51	94.12	91.78	90.50
6 23 145	92.09	95.29	93.74	92.31
6 24 135	88.99	94.09	91.20	90.14
16 25 34	88.79	93.94	90.70	89.89
16 23 45	91.89	95.24	93.93	92.25
6 15 234	89.14	94.20	90.64	90.15
4 16 235	88.87	93.87	91.09	89.73
14 25 36	88.55	93.96	91.05	89.49
6 13 245	92.30	95.58	93.87	92.92
12 36 45	92.07	95.39	93.87	92.47
6 45 123	92.30	95.46	94.04	92.78
2345 1 6	92.75	95.66	93.78	93.48
16 24 35	88.88	94.18	89.76	90.16
1 25 346	90.09	94.94	91.74	90.70
6 14 235	88.83	93.86	89.99	89.86
4 36 125	88.97	93.99	90.53	89.74
6 34 125	89.28	94.28	90.75	90.10
1245 3 6	92.85	95.54	93.63	93.28
1345 2 6	92.20	95.47	93.09	92.94
2 16 345	91.96	95.42	93.57	92.55
2346 1 5	90.32	95.27	91.48	90.92

Fonte: Elaborada pelo autor (2019)

Assim, o melhor agrupamento é 2 13 456, que possui o melhor *trade-off* entre acurácia, sensibilidade e especificidade com uma acurácia de 99.23% nas Redes Neurais. É possível observar ainda que as melhores acurácias estão relacionadas à presença dos itens de inspeção 4, 5 e 6 em um mesmo grupo. Assim, esses três itens poderiam ser agrupados para indicação de um nível de saúde. Bem como os itens 1 e 3. Portanto, um alerta poderia ser emitido caso qualquer item do grupo 456 ou 13 apresente um problema.

Vale destacar que os itens 4, 5 e 6 da planilha de inspeção são itens que não estão diretamente ligado à colônia em si: alimento, estressores e espaço, respectivamente. O agrupamento formado pelos itens 1 e 3 possui os itens da planilha de inspeção diretamente ligados à rainha. Uma vez que o item 1 (cria) representa todas as fases da cria e ínstares.

4.3 Sumário do Capítulo

Como na maioria dos estudos, devemos reconhecer as limitações e advertências associadas a esta pesquisa. Primeiro, os apiários dos quais os pontos amostrais foram retirados para o estudo estão todos localizados nos Estados Unidos da América, possivelmente limitando a generalização. A influência da variedade de abelhas e variações climáticas encontradas fora dos EUA precisará ser investigada em estudos futuros. Da mesma forma, a diferença nas técnicas de apicultura e inspeção pode influenciar a coleta de dados do estudo. No entanto, dentro do estudo, foram tomadas medidas para minimizar possíveis vieses. Todos os apicultores participantes do estudo foram treinados no uso da lista de verificação por seu autor original e seguiram práticas semelhantes de apicultura, utilizando os mesmos tipos de equipamento.

Finalmente, a categorização do status é de natureza transversal. Em outras palavras, a categorização é para o momento em que os dados do sensor foram obtidos. A previsão do futuro estado de saúde de uma colônia deve ser abordada com cautela, pois o desenho do estudo não incorpora as facetas temporais dos dados.

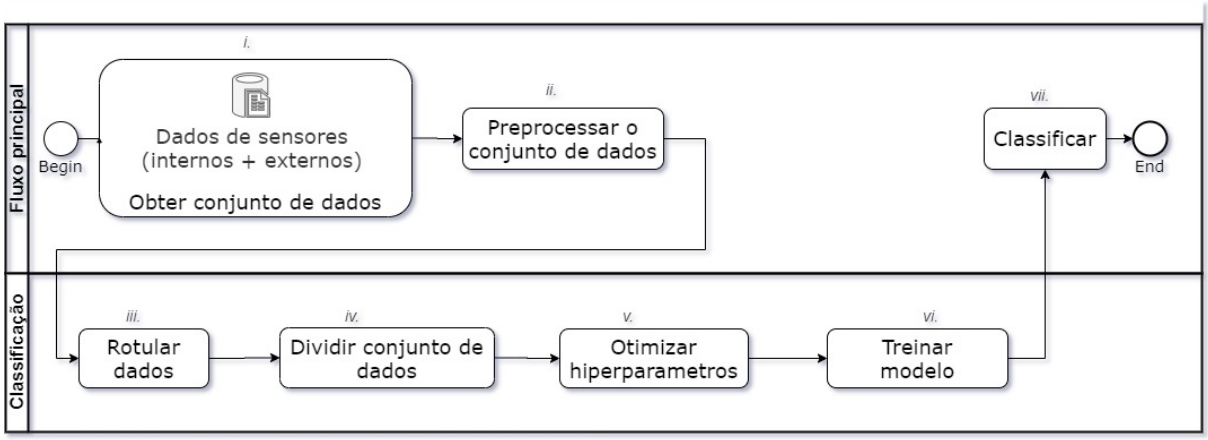
5 REGRESSÃO

Este capítulo descreve os aspectos metodológicos e os resultados da abordagem que utiliza a técnica de mireação de dados regressão, detalhando tópicos relacionados às ferramentas utilizadas, coleta e pré-processamento de dados, bem como a análise e detecção dos níveis de bem estar de colônia de abelhas utilizando a regressão. Para a validação dessa abordagem, foi utilizado um conjunto de dados específico para essa abordagem conforme recomendado no Capítulo 1.

Nessa abordagem, os possíveis níveis de bem estar são definidos previamente, em termos quantitativos e qualitativos. As etapas a serem realizadas para obteção do modelo de classificação são ilustradas na Figura 30. A Seção 5.1 detalha cada uma dessas etapas. Lembrando que o processo ou método da Figura 30 é uma instância do método KDD apresentado no Capítulo 2.

Na regressão, os níveis de bem estar foram preditos de forma indireta através da identificação da perda da capacidade de termorregulação da colônia. A predição da perda da termorregulação é realizada através de um algoritmos de regressão treinado, validado e testado.

Figura 30 – Método de predição via aprendizado supervisionado (regressão)



Fonte: Elaborada pelo autor (2019)

5.1 Materiais e Métodos

Para validar o método aplicado de predição da perda da termoregulação em colmeias baseado em regressão, foram usados dados do software de monitoramento de colmeias Arnia[1]. O

[1] http://www.arnia.co.uk/

Arnia é um sistema de monitoramenoto de colmeias fundado na Inglaterra em 2009. Atualmente, o sistemas está funcionando em mais de 25 países.

5.1.1 Conjunto de dados

O conjunto de dados utilizado é composto de dados de cinco colmeias e foi obtido de um apiário localizado na cidade de Newcastle upon Tyne, Inglaterra, distante 447 km do norte de Londres. Eles foram coletados de setembro a novembro de 2017, durante o outono. O conjunto de dados possui 6 atributos; (i) temperatura das crias (temperatura interna), (ii) umidade interna, (iii) ventilação média, (iv) ruído médio de vôo, (v) peso e (vi) temperatura externa. As seis características foram utilizadas para prever a temperatura interna de cada colmeia usando séries temporais. A temperatura interna da colmeia e a temperatura externa foram medidas em graus Celsius. A Tabela 14 apresenta um resumo da amostragem realizada para cada colmeia, incluindo o número de pontos amostrais para cada colmeia. Na coluna "term." (isto é, termorregulada), b.r. significa *bem regulado* e n.r. significa *não regulado*.

Tabela 14 – Resumo das colmeias analisadas

Colmeias	Latitude	Longitude	#amostras	Período	Amostragem	Term.
9803	-1.628	54.971	603	Sep 3th-Nov 6th	2hs	b.r.
9841	-1.617	54.979	638	Sep 5th-Nov 6th	2hs	n.r.
9848	-1.599	55.016	502	Sep 5th-Nov 2th	2hs	n.r.
54440	-1.628	54.971	606	Sep 3th-Nov 6th	2hs	n.r.
54460	-1.616	54.970	1024	Aug 5th-Nov 6th	2hs	b.r.

Fonte: Elaborada pelo autor (2019)

O sistema Arnia usa três sensores dentro da colmeia para coletar os atributos que descrevem as condições internas da colmeia. Um sensor de temperatura, um sensor de umidade, uma balança digital e um sensor de som (microfone). Os atributos podem ser agrupados por objetivo: temperatura/umidade das crias (homeostase da colmeia), voo/forrageamento e acústica da ventilação (atividade das abelhas) e peso da colmeia (produtividade). As colônias em estudo são da espécie *Apis mellifera*.

5.1.2 Preprocessamento

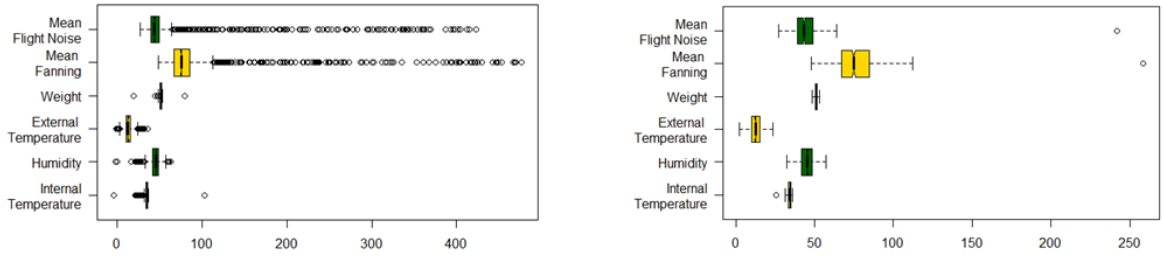
Inicialmente, foi realizada a Análise Exploratória de Dados, do inglês *Exploratory Data Analysis* (EDA), utilizando métodos estatísticos, para familiarizar-se com os dados e maximizar a recuperação de informações ocultas em sua estrutura. Depois disso, foram calculadas

estatísticas básicas (média, desvio padrão e quartis) e a assimetria. Em seguida, foram desenhados os gráficos de dispersão e histogramas para ver a distribuição dos dados. Por fim, realizamos a detecção e remoção de valores discrepantes e o redimensionamento dos dados definidos pelo método min-max.

5.1.2.1 Detecção e remoção de outliers

O boxplot plotado de cada colmeia mostrou que havia alguns *outliers* em todos os atributos utilizados, um exemplo disso pode ser visto na Figura 31(a) para a colmeia 54460. Para detectar os valores discrepantes, foi utilizado o cálculo do intervalo interquartil, que é a diferença entre os quartis 1 e 3 do conjunto de dados. Os números que estavam nos quartis fora desse intervalo foram transformados em NULL. O método de imputação escolhido foi da média, que substitui o valor ausente pela média do recurso. Após essa remoção, o resultado é um boxplot com apenas alguns outliers exemplificados na Figura 31(b).

Figura 31 – Diagramas de caixas dos atributos da colmeia 54460



(a) Com os *outliers* (b) Sem os *outliers*

Fonte: Elaboradas pelo autor (2019)

5.1.2.2 Redimencionamento de dados (reescalonamento min-max)

Como o LSTM é sensíveis à escala dos dados de entrada, os dados foram redimensionados para que a rede neural os interprete adequadamente. Além disso, isso geralmente é usado para melhorar a estabilidade numérica de um modelo. A técnica min-max pega os valores dos atributos em seus intervalos originais e os redimensiona para um intervalo de [0,1]. Isso pode ser útil em alguns casos em que todos os parâmetros precisam ter a mesma escala positiva. O

conjunto de dados foi normalizado pela Equação 5.1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{5.1}$$

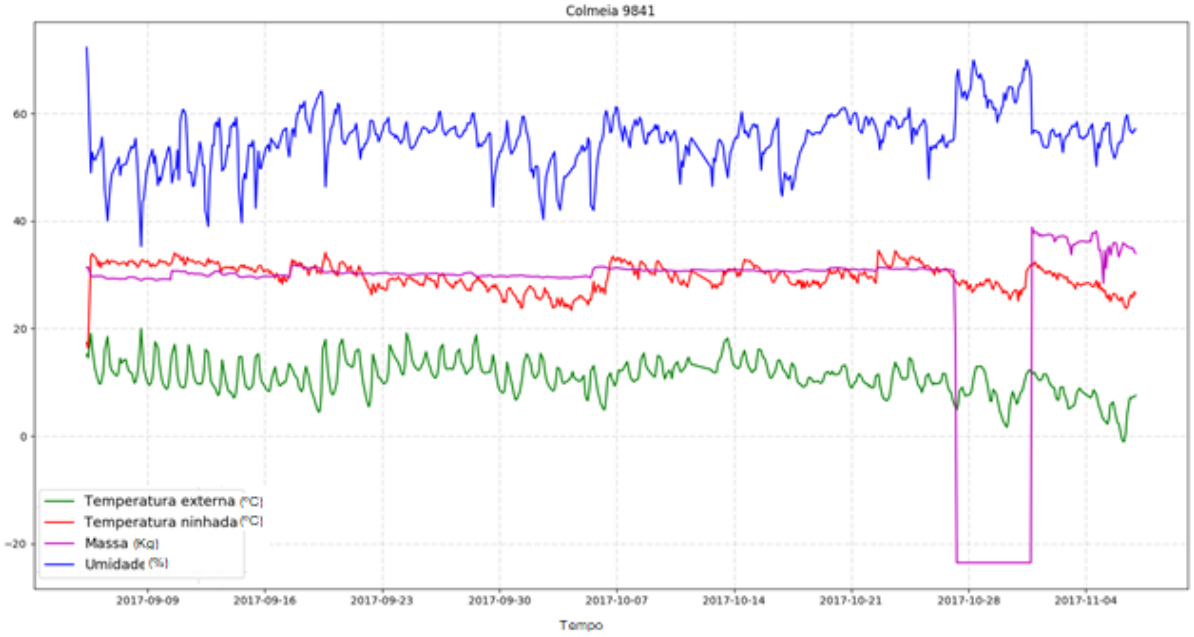
5.1.3 Rotulagem do conjunto de dados

A rotulagem buscou a caracterização de séries temporais segundo a capacidade de manutenção da termorregulação nas colmeias. O processo se deu através de entrevista com um especialista em apicultura. Durante a entrevista, gráficos de linha no tempo das variáveis em estudo foram apresentadas para o especialista. Os dados, que estão dispersos em longo de 2 ou 3 meses, entre Agosto e Novembro, foram divididos em janelas de tempo de 1 semana para apresentação. Assim, o especialista foi capaz de identificar as semanas em que uma determinada colônia possuía ou não um bom controle de termorregulação. A Tabela 14 apresenta na coluna "term." as colmeias que apresentaram bom controle termorregulatório (b.r) ou não (n.r). A Figuras 32, 33, 34, 35 e 36 mostram os graficos de linha mencionados acima.

Figura 32 – Gráfico de linha da colméia 9803

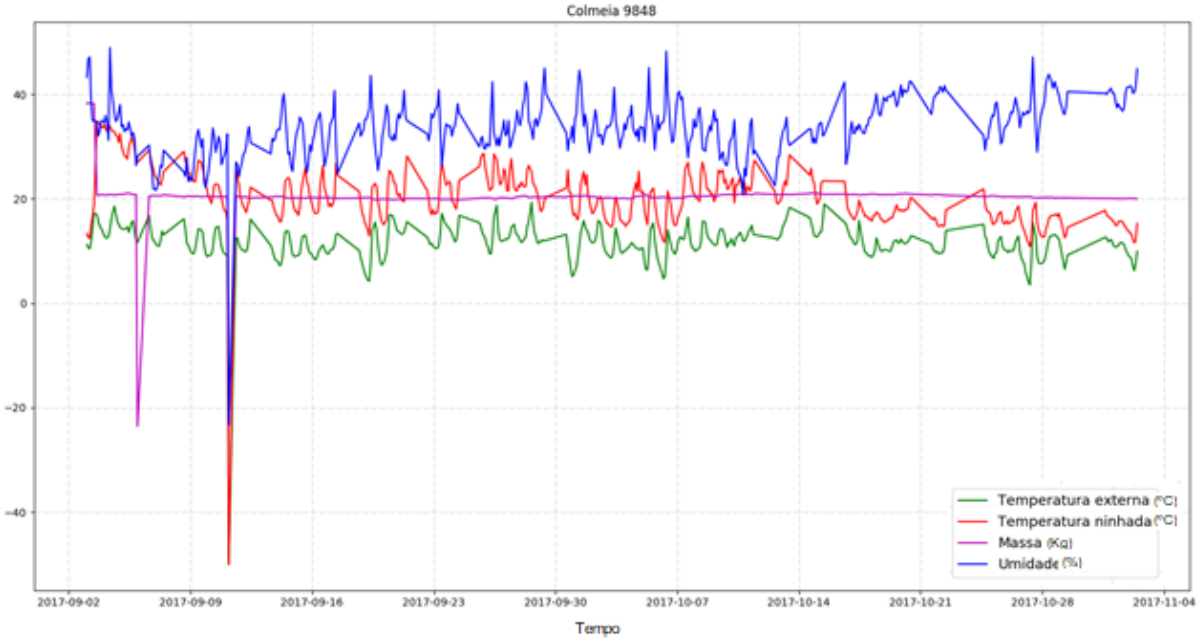
Fonte: Elaborada pelo autor (2019)

Figura 33 – Gráfico de linha da colméia 9841



Fonte: Elaborada pelo autor (2019)

Figura 34 – Gráfico de linha da colméia 9848



Fonte: Elaborada pelo autor (2019)

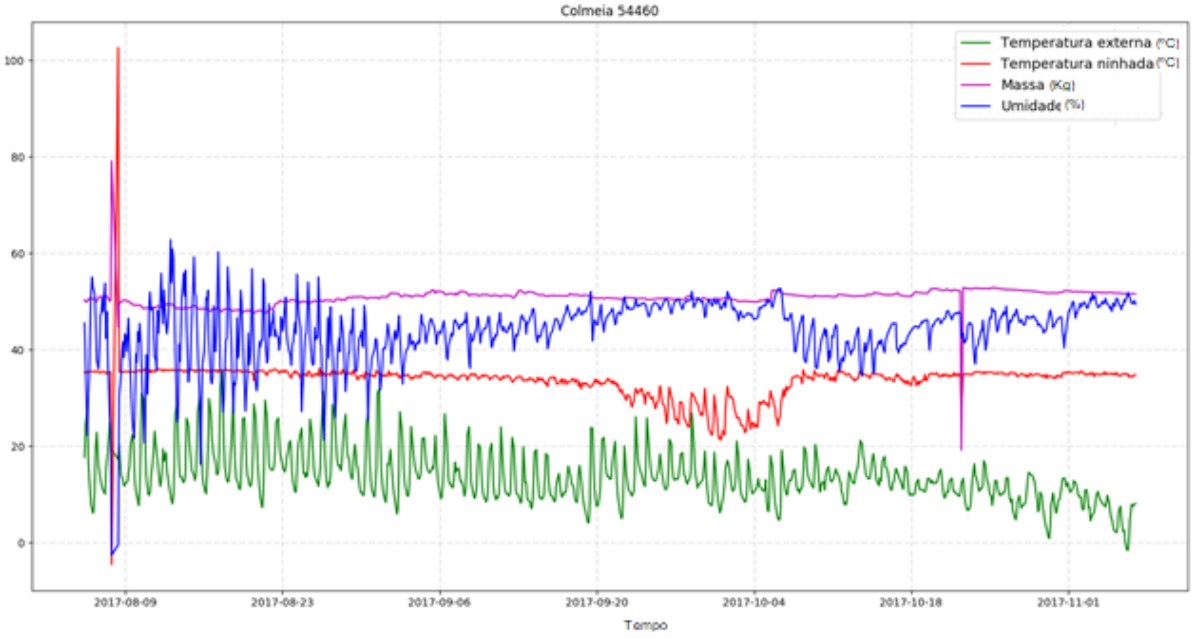
5.1.4 Divisão do conjunto de dados e validação walk-forward

O conjunto de dados de cada colmeia foi dividido na seguinte proporção: 50% para treinar, 20% de conjunto para validação e 30% do conjunto para testes. A validação foi usada para ajustar os parâmetros, que serão discutidos em mais detalhes na Seção 5.1.5.

Figura 35 – Gráfico de linha da colméia 54440

Fonte: Elaborada pelo autor (2019)

Figura 36 – Gráfico de linha da colméia 54460

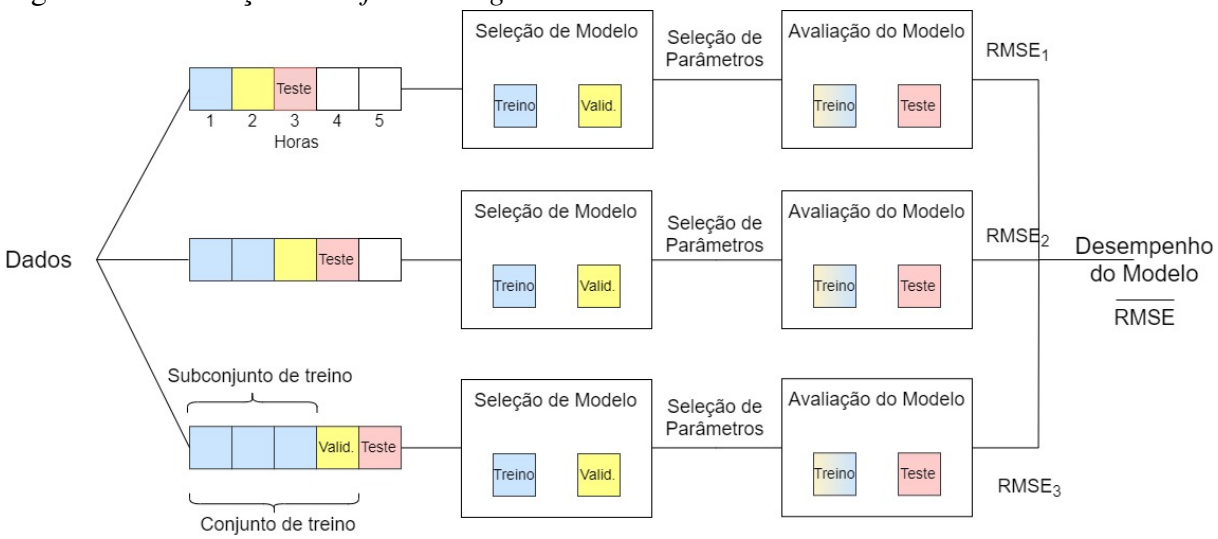


Fonte: Elaborada pelo autor (2019)

Para avaliar os modelos de uma maneira mais precisa, foi usado um processo chamado validação *walk-forward* (HU *et al.*, 1999), conhecido como validação cruzada ou *k-fold cross validation* dos problemas de séries temporais. Primeiro, para esse método, um tamanho de janela deslizante é escolhido para treinar o conjunto de dados e, em seguida, o modelo faz uma previsão na próxima etapa (Figura 37). Essa previsão é armazenada e analisada com o valor

real do conjunto de dados, para que possamos ter métricas como o *Root Mean Squared Error* (RMSE) que calcula a diferença entre esses dois valores. Esse processo acontece novamente com a janela deslizando para dados que não foram usados anteriormente até o final do conjunto de dados.

Figura 37 – Validação *walk-forwarding*



Fonte: Elaborada pelo autor (2019)

A validação *walk-forwarding* tem o benefício de fornecer uma estimativa muito mais realista de como o método e os parâmetros de modelagem escolhidos serão executados na prática.

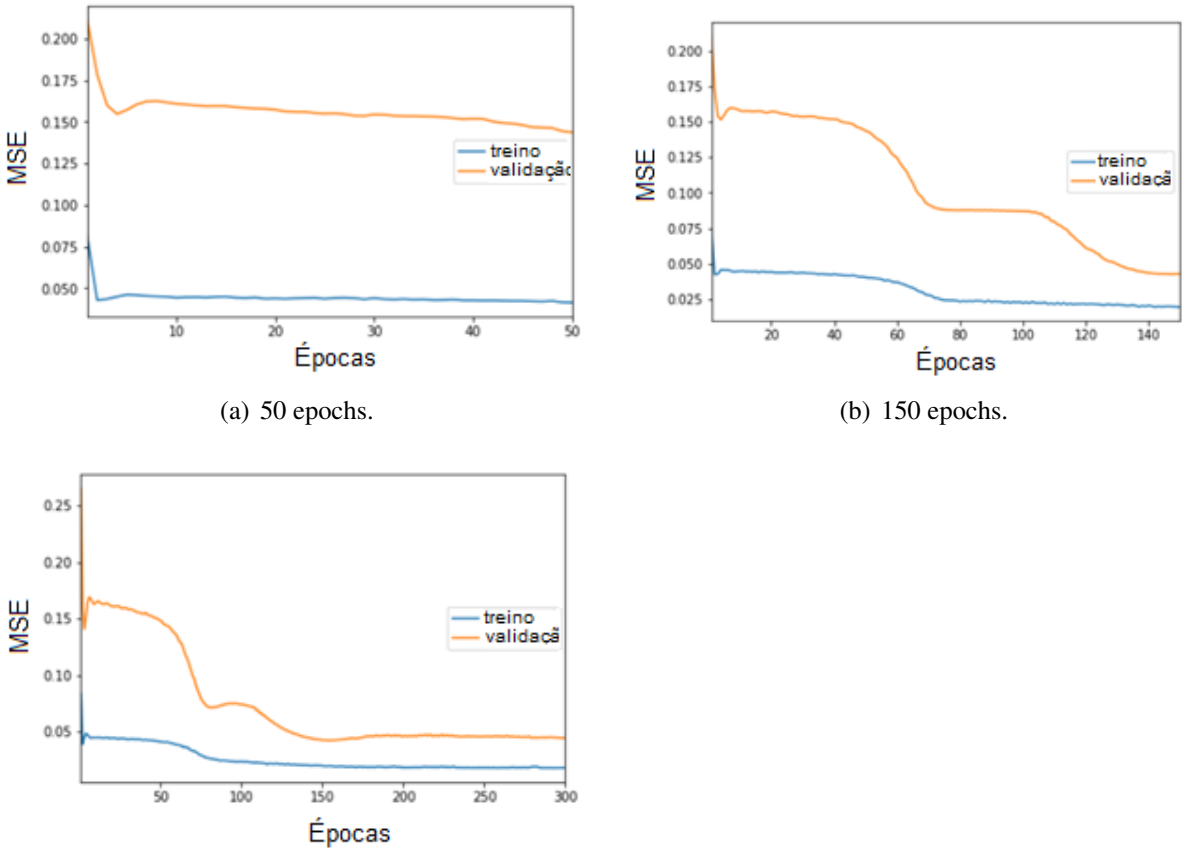
5.1.5 Otimização dos hiperparâmetros

Como as redes neurais podem criar modelos realmente complexos, existem vários parâmetros que precisam ser ajustados para alcançar o melhor modelo para o problema discutido. Por exemplo, o número de vezes que o conjunto de dados é treinado na rede neural, as iterações, também chamadas de épocas, pode ser alcançado através da experimentação. Inicialmente, foi experimentado uma quantidade de 50 épocas. Para essa configuração, os erros de treinamento e validação foram examinados, como mostra a Figura 38(a). O erro de validação apresentou-se alto em comparação ao erro do conjunto de treino, além disso, não estavam convergindo. Isso indicou um *underfit* no modelo, ou seja, um bom desempenho no conjunto de treinamento e baixo desempenho no conjunto de validação. Assim, por experimentação, resolveu-se aumentar o número de épocas para 150. O resultado foi consideravelmente melhor, vide Figura 38(b). Contudo, pode-se notar que o erro de validação mostrou o potencial de diminuir ainda mais.

Dessa forma, resolveu-se aumentar a quantidade de épocas para 300. Conforme representado na Figura 38(c), os erros convergiram e não havia nenhum sinal de *underfitting* ou *overfitting*. O tamanho do lote (*batch*) selecionado foi o padrão, 32, pois, quando ajustado, não provou ter uma grande diferença no resultado.

Para configurar o processo de aprendizagem, dois parâmetros importantes foram definidos. O otimizador, que é um algoritmo de otimização que minimiza a "função de erro". Para o modelo aplicado, o otimizador escolhido foi o AdamX, um algoritmo de otimização baseado em um gradiente de primeira ordem de funções objetivas estocásticas, que é uma variante de Adam baseada na norma infinita. Além disso, como função de perda, escolhemos o Erro Quadrático Médio, do inglês *Mean Square Error* (MSE), cujo o objetivo do modelo e do otimizador é minimizá-la. A métrica MSE é detalhada no Anexo B.2.

Figura 38 – MSE para a colmeia 9803 usando uma janela deslizante de 24 horas.



(a) 50 epochs.

(b) 150 epochs.

(c) 300 epochs.

Fonte: Elaboradas pelo autor (2019)

5.1.5.1 Definição da arquitetura do modelo

A arquitetura da rede LSTM foi definida de maneira experimental, ou seja, adicionando-se camadas na rede neural para criar um modelo ideal. Começamos com a arquitetura mais simples composta por uma camada LSTM com 50 camadas ocultas (*hidden layers*) e uma camada Densa (*Dense Layer*) com 1 camada oculta (**arquitetura #1**). Na Figura 39(a) é possível verificar a aplicação da arquitetura # 1 na colmeia 9848.

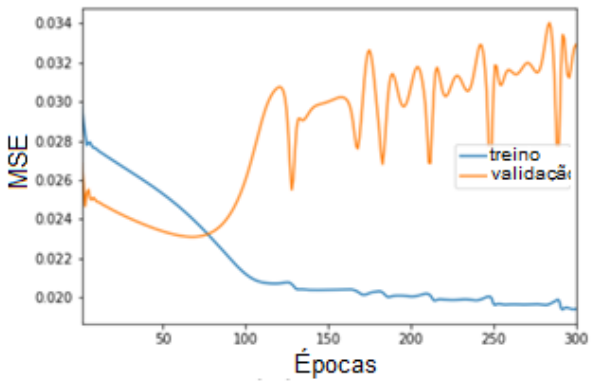
O erro de validação foi aumentando com as épocas, devido a essa natureza de *overfitting* excessiva. Portanto, devido ao número de pontos amostrais e à complexidade do problema, essa configuração não conseguiu predizer com um bom resultado. Em seguida, optamos por adicionar uma camada Dropout com uma taxa de 0,5 unidades de entrada a serem removidas (**arquitetura #2** - ou seja, uma camada LSTM com 50 camadas ocultas, uma camada Dropout de 50% e uma camada densa com uma camada oculta). O erro de validação começou a se estabilizar como visto na Figura 39(b). Analisando todo o conjunto de dados das colmeias, foi possível observar que era necessário aumentar a capacidade do modelo para fazer convergir o erro de treinamento e validação. Por esse motivo, aplicamos mais uma camada Densa com 50 camadas ocultas (**arquitetura #3** - uma camada LSTM com 50 camadas ocultas, uma camada Dropout de 50%, Camada Densa com 50 camadas ocultas e uma Camada Densa com 1 camada oculta). Como visto na Figura 39(c) e 39(d), o erro de validação foi baixo e próximo ao erro de treinamento nas colmeias 9841 e 9803. Por fim, para evitar *overfitting* adicionais, aplicamos outra camada Dropout com uma taxa de 0,5, ou seja, **arquitetura #4** - uma camada LSTM com 50 camadas ocultas, 2 camada Dropout de 50%, camada densa com 50 camadas ocultas, e uma camada densa com 1 camada oculta. Isso melhorou os resultados de colmeias mais complicadas de prever, como 9848, Figura 39(e), e não afetou os bons erros de validação das outras colmeias.

5.2 Resultados e Discussões

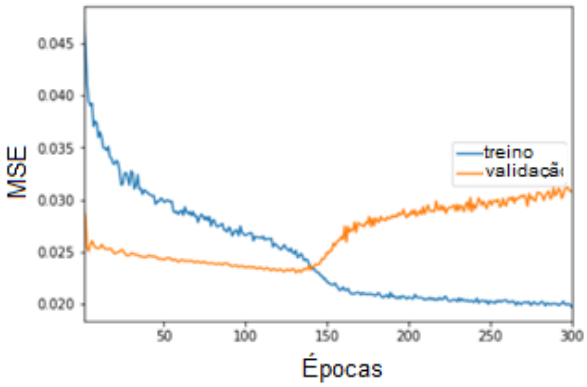
Como escolheu-se a validação *walk-forward*, um tamanho de janela deslizante precisou ser escolhido para realizar os experimentos (GOMES *et al.*, 2020). Essa janela representa quantas horas o algoritmo pode prever com antecedência a temperatura interna. Foram testados três possíveis tamanhos de janelas deslizantes: 2, 10 e 24 horas.

O modelo final alcançado foi um LSTM com 100 camadas ocultas, 2 camadas Dropout de 50%, uma camada Densa com 50 camadas ocultas e, finalmente, uma camada 1

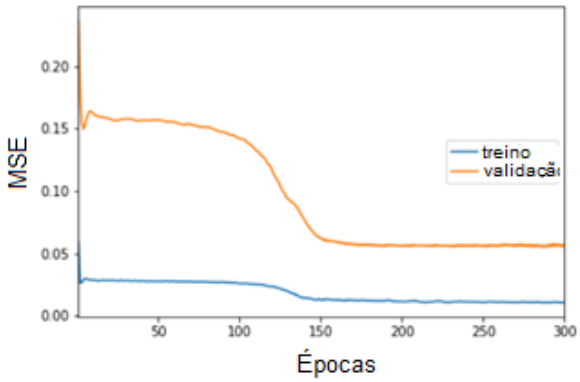
Figura 39 – Seleção da arquitetura do modelo



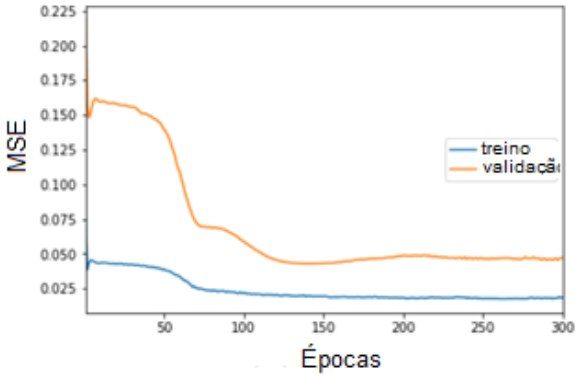
(a) Arquitetura #1 na colmeia 9848.



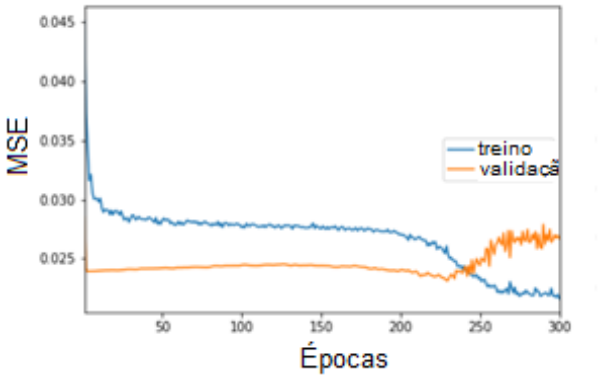
(b) Arquitetura #2 na colmeia 9848.



(c) Arquitetura #3 na colmeia 9841.



(d) Arquitetura #3 na colmeia 9803.



(e) Arquitetura #4 na colmeia 9848.

Fonte: Elaboradas pelo autor (2019)

Densa. Todos os conjuntos de dados foram treinados com 300 épocas e um tamanho de lote de 32. Além disso, outro parâmetro foram definido: o número de horas que o algoritmo usaria para prever a saída, janela de entrada. Seguiu-se duas abordagens, a primeira na qual o número de horas da janela de entrada era o dobro da janela deslizante e a segunda onde a janela de entrada era o mesmo número da janela deslizante.

Para avaliar a qualidade dos modelos aplicados, foi utilizada a métrica RMSE. O

RMSE mede a média dos quadrados dos erros, ou seja, a diferença quadrática média entre os valores estimados e o que é estimado. O RMSE possui valores baixos para um bom modelo. Os resultados podem ser vistos na Tabela 15.

Tabela 15 – RMSE de teste para cada colmeia

Experimento/colmeia	**9803**	**9841**	**9848**	**54440**	**54460**
janela deslizante = 2 horas janela de entrada = 2 horas	0.215	0.154	0.171	0.217	0.042
janela deslizante = 2 horas janela de entrada = 4 horas	0.149	0.153	0.125	0.153	0.054
janela deslizante = 10 horas janela de entrada = 10 horas	0.233	0.138	0.193	0.227	0.067
janela deslizante = 10 horas janela de entrada = 20 horas	0.203	0.146	0.184	0.207	0.053
janela deslizante = 24 horas janela de entrada = 24 horas	0.230	0.150	0.156	0.227	0.056
janela deslizante = 24 horas janela de entrada = 48 horas	0.246	0.199	0.166	0.250	0.061

Fonte: Elaborada pelo autor (2019)

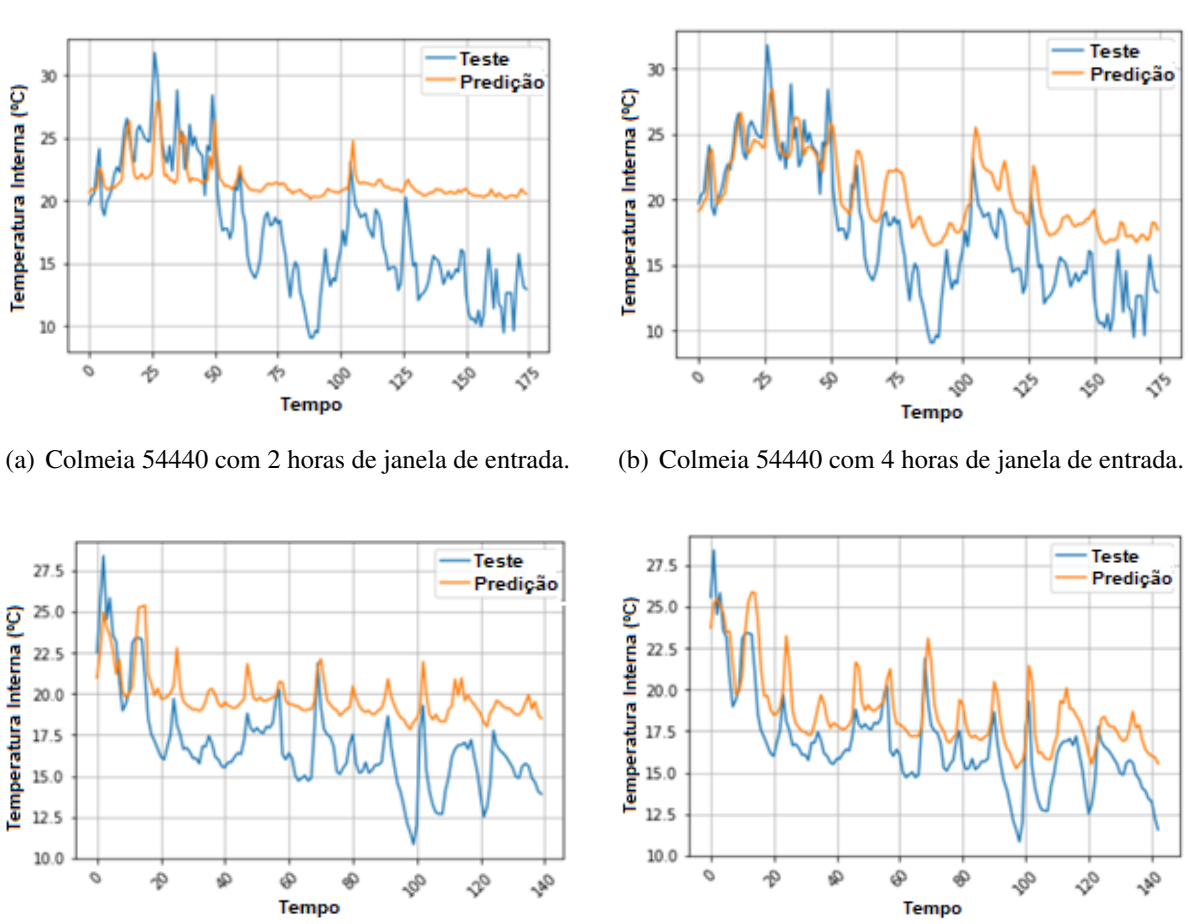
5.2.0.1 Previsão com 2 horas de antecedência

Nesse experimento foi estabelecida uma janela deslizante de 2 horas de antecedência para prever a temperatura interna. A primeira parte do experimento foi com uma janela de entrada de 2 horas. Para visualizar melhor a saída da previsão e entender os valores de RMSE, apresenta-se as Figuras 40(a) e 40(c), das colmeias 54440 e 9848, respectivamente, os gráficos com os valores de teste e preditos. Na segunda parte do experimento, a janela de entrada foi de 4 horas com o mesmo intervalo de previsão (2 horas). Já pode-se notar uma diferença nos resultados observando a Figura 40(b) e 40(d), que mostra os gráficos das mesmas colmeias (54440 e 9848, respectivamente) para esta nova configuração.

5.2.0.2 Previsão com 10 horas de antecedência

No experimento a seguir, foi utilizada uma janela deslizante de 10 horas de antecedência para prever a temperatura interna. Em relação da janela de entrada, o algoritmo LSTM foi treinado com janelas de 10 e 20 horas. Os RMSE's de teste e de predição obtidos estão descritos na Tabela15. Para a colmeia 54460, o RMSE foi de 6.7% para a janela de entrada de 10 horas

Figura 40 – Predição da temperatura interna com janela deslizante de 2 horas



(a) Colmeia 54440 com 2 horas de janela de entrada.

(b) Colmeia 54440 com 4 horas de janela de entrada.

(c) Colmeia 9848 com 2 horas de janela de entrada.

(d) Colmeia 9848 com 4 horas de janela de entrada.

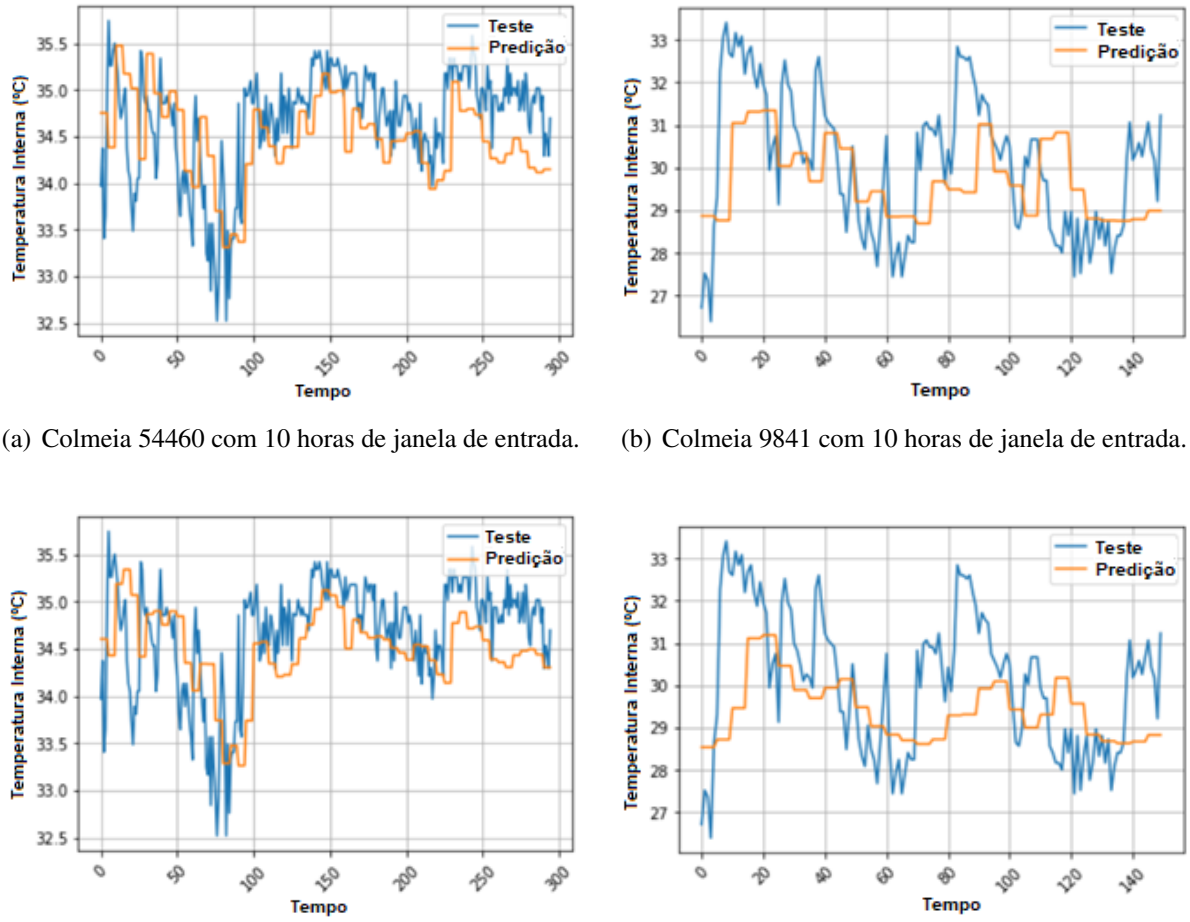
Fonte: Elaboradas pelos autor (2019)

e 5.3% quando a janela de entrada foi de 20 horas. Nos gráficos mostrados nas Figura 41(a) e 41(c), podemos ver as tendências da linha de previsão na temperatura seguindo a forma no conjunto de dados de teste. No caso da colmeia 9841, o RMSE foi maior com 13.8% e 14.6% para as janelas de entrada de 10 e 20 horas, respectivamente. Devido ao erro ser maior, pode-se ver nos gráficos das figuras 41(b) e 41(d) que a temperatura prevista tem valores maiores que os da temperatura de o conjunto de dados, mas ainda segue bem o padrão.

5.2.0.3 Previsão com 24 horas de antecedência

Neste experirimento foi testada uma janela deslizante de 24 horas de antecedência para prever a temperatura interna. Os RMSE's de teste também estão na Tabela 15 e mostra as duas abordagens, com as janelas de entrada de 24 horas e de 48 horas. Observando os valores obtidos, para as colmeias 9841 e 9803, é possível observar os valores so RMSE de 19,9% e

Figura 41 – Predição da temperatura interna com janela deslizante de 10 horas



(a) Colmeia 54460 com 10 horas de janela de entrada.

(b) Colmeia 9841 com 10 horas de janela de entrada.

(c) Colmeia 54460 com 20 horas de janela de entrada.

(d) Colmeia 9841 com 20 horas de janela de entrada.

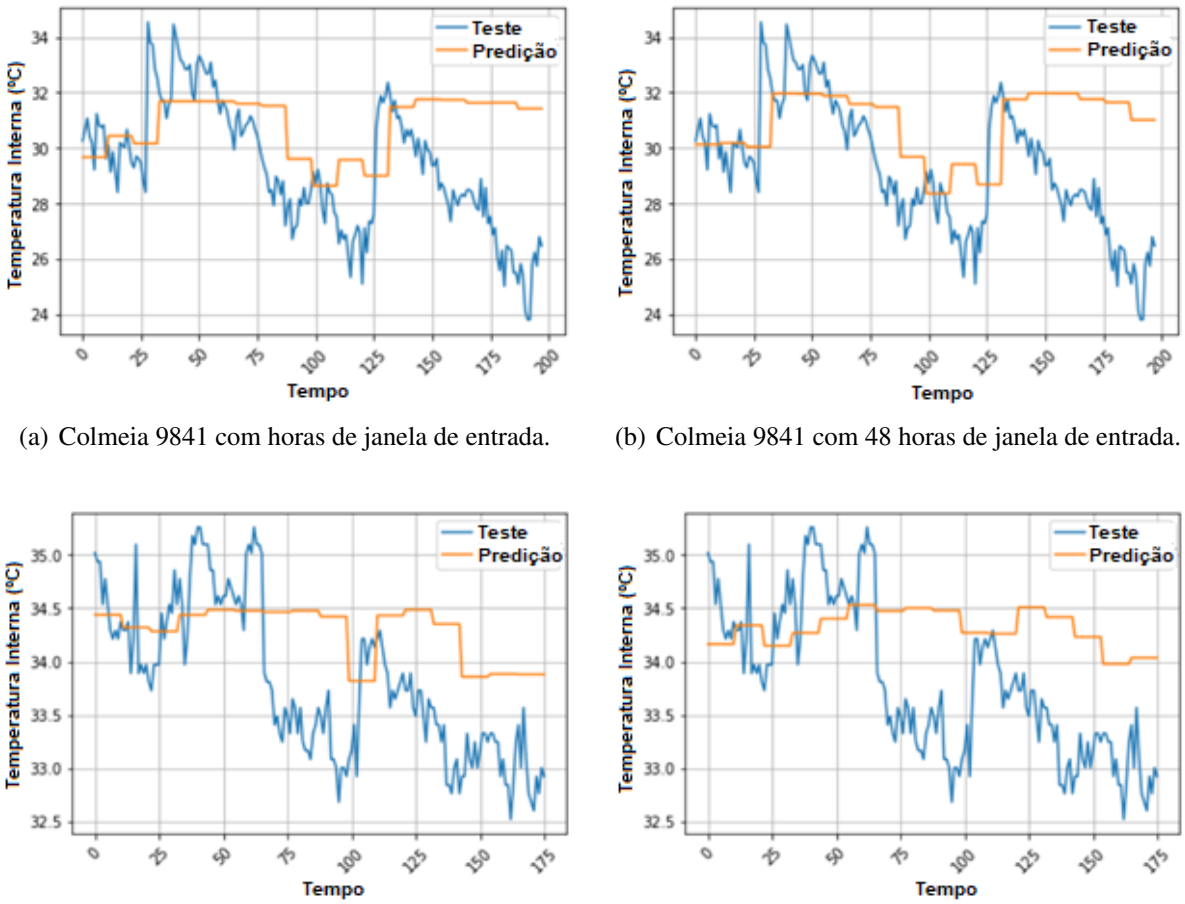
Fonte: Elaboradas pelo autor (2019)

24,6%, respectivamente, que são erros expressivos. Os erros já mostram que o modelo não fornece resultados tão satisfatórios, mas analisando os gráficos das Figuras 42(a) e 42(d), é possível observar que a temperatura prevista não segue bem a temperatura de teste para as duas colmeias analisadas.

Para as colmeias 9803, 9848, 54460, 54440, o RMSE calculado para a predição no primeiro experimento foi o mais baixo comparado com os obtidos nos outros experimentos. Analisando essas colmeias, notou-se uma chance de erros menores com uma janela de entrada de 2 horas, o que representa um modelo que pode prever melhor a temperatura interna da colmeia. Diante disso, foi possível observar que, quando a janela de entrada é duas vezes maior que a janela de previsão, os resultados são melhores, isso pode ser visto visualmente comparando as Figuras 40(a) e 40(b).

No segundo experimento, o RMSE foi baixo em geral, mas tiveram um aumento

Figura 42 – Predição da temperatura interna com janela deslizante de 24 horas



(a) Colmeia 9841 com horas de janela de entrada.

(b) Colmeia 9841 com 48 horas de janela de entrada.

(c) Colmeia 9803 com 24 horas de janela de entrada.

(d) Colmeia 9803 com 48 horas de janela de entrada.

Fonte: Elaboradas pelo autor (2019)

considerável em comparação aos resultados do experimento 1, especialmente para as colmeias 9848 e 9803. Nesse experimento, optou-se por aumentar a janela de entrada por 20 horas para ver se, com mais pontos amostrais, os resultados dos testes melhoraria, foi possível observar melhorias para as colmeias 9803, 9848, 54440 e 54460, onde o RMSE diminuiu.

No terceiro experimento, como mostrado na Figura 42(c), o modelo não conseguiu prever os padrões de temperatura. Mesmo quando, na tentativa de melhorar esses resultados, foi realizada uma configuração com uma janela de entrada de 48 horas, esse cenário não foi alterado. De fato, para as colmeias 9803, 9841 e 54440, obteve-se o pior RMSE e os gráficos mostraram incapacidade de seguir o padrão de temperatura.

Assim, foi possível mostrar que o algoritmo LSTM foi capaz de prever a amplitude e os valores da temperatura interna das colmeias. Foi possível também avaliar o estado das colmeias quanto a capacidade de manter termorregulação ou não. A seguir, será feita uma discussão dos resultados obtidos.

A alta amplitude de temperatura e um padrão em declínio indicam um sinal de perda da capacidade de termorregulação de uma colônia. Nos conjuntos de dados utilizados nessa experimentação, a estação em que os dados foram coletados por meio de sensoriamento foi o outono. Durante o outono, a temperatura cai gradualmente, especialmente à noite, em preparação para o inverno. Portanto, esperava-se que a temperatura interna também caísse de acordo com a temperatura externa (WINSTON, 1991). No entanto, dado que nas colmeias 9841, 9848 e 54440 a temperatura caiu mais que 10°C, isso pode ser um sinal da perda da capacidade de termorregulação, o que poderia levar a vários problemas no bem estar da colônia (HELDMAIER, 1987).

Além disso, nesse trabalho trabalho, mostramos colmeias onde as colônias de abelhas conseguiram termorregular bem e nosso algoritmo também pode prever esse padrão, no caso as colmeias 9803 e 54460. Um controle adequado da temperatura é crucial para a sobrevivência das abelhas e a reprodução nesta espécie. Os efeitos de nossos resultados são importantes para o trabalho diário do apicultor, pois ele poderá preparar ou até evitar esse problema antes que ele aconteça. Além disso, o apicultor poderá adaptar o manejo das colônias para prepará-las, melhorando seu microclima de acordo com os resultados fornecidos pela previsão.

A aplicação do LSTM pode ser considerada também para predição de outros fenômenos importantes relacionados com o ciclo de vida das colônias de abelhas, tais como: enxameamento, perda da rainha, abandono, distúrbio do colapso das colônias, do inglês, *Colony Collapse Disorder* (CCD) e fluxo de nectar. Além do LSTM, que faz uma regressão, outros tipos de estratégias de aprendizado de máquina/algoritmos poderiam ser utilizados com uma promissora eficiência, tais como: o agrupamento ou classificação de séries temporais e os modelo auto-regressivo integrado de médias móveis, do inglês, *Auto-Regressive Integrated Moving Average* (ARIMA). O LSTM apresentou um baixo erro provavelmente devido a sua arquitetura de células para armazenamento de infomações e, consequentemente, diminuição do erro.

5.3 Sumário do Capítulo

As limitações e advertências associadas ao uso desse método são destacadas abaixo. Primeiro, as colônias usadas neste estudo estão todas localizadas no Reino Unido, possivelmente restringindo a generalização. A influência da variedade de abelhas e variações climáticas encontradas fora do Reino Unido precisará ser investigada em estudos futuros. Da mesma forma, a diferença no especialista em apicultura pode influenciar a rotulação das séries temporais

neste método. No entanto, foram tomadas medidas para minimizar possíveis viés através da orientação do especialista para padronizar o processo de rotulagem. Portanto, recomenda-se que o especialista em rotulagem leve em consideração: a estatística descritiva de cada colmeia (incluindo média, mediana, modo, desvio padrão, variância, valor máximo e mínimo, obliquidade e curtose), o período do ano (estação) , a região geográfica e a espécie da abelha em estudo.

6 CONCLUSÃO

Neste capítulo de conclusão, as três questões de pesquisa levantadas na Introdução desta tese são respondidas, suas respectivas hipóteses são confirmadas e são apresentadas também perspectivas de trabalhos futuros para cada método. A Seção 6.4 lista as publicações realizadas.

6.1 Agrupamento - QP #1

Essa tese apresentou uma solução baseada nas técnicas de mineração de dados agrupamento e classificação para detectar padrões sazonais de colônias de abelhas que pode ser personalizado e integrado a um sistema de computador de monitoramento de apiários para recomendação de manejos proativos a fim de evitar perdas de colônias (**Contribuição #1**). A literatura geralmente considera apenas uma variável física, geralmente temperatura, para monitoramento e análise; em alguns casos, outras variáveis como por exemplo umidade, mas sem abordar a análise de dados monitorada. O método aqui aplicado leva em consideração cinco variáveis (temperatura interna e externa, umidade interna e externa e peso da colmeia), bem como suas respectivas análises e identificação dos valores típicos de uma colônia forte ou fraca anualmente e entre as estações (resposta para a **QP #1**). Com taxas de acerto de até **99,67**%, o método aplicado detectou 6 padrões reconhecidos como coerentes, pois correspondem ao que é observado em campo. As principais conclusões indicam que a colônia mais forte é a mais eficiente na manutenção da estabilidade do microclima da colmeia durante o inverno. Portanto, o método baseado em agrupamento mitiga o **problema** apresentado na Seção 1.2 e confirma a **Hipótese #1** (Seção 1.3.1).

Como estudos futuros a curto e médio prazo, planejamos aplicar nosso método em conjuntos de dados adicionais, por exemplo: (i) dados de colônias dos EUA do sistema hivetools, (ii) dados de colônias de abelhas africanizadas (*Apis mellifera*) no Brasil e (iii) dados de colônias do projeto BBCC, o mesmo utilizado para validar a solução baseada apenas em classificação. Para gerar os dados de abelhas africanizadas, um sistema de monitoramento remoto, Sm@rtBee [1] está sendo desenvolvido. Em termos das técnicas de mineração de dados, uma possibilidade de investigação futura é a utilização da evolução de grupos (SPILIOPOULOU *et al.*, 2006; SILVA *et al.*, 2014). Com essa abordagem, seria possível observar a transformação dos grupos ao longo do tempo e, consequentemente, prever com maior precisão quando uma colmeia estaria por

[1] http://smartbee.great.ufc.br/

entrar em um estado não desejado. Outra possível vantagem dessa técnica seria a observação mais detalhada dos grupos e suas variações, permitindo a identificação de outros padrões não observados com a abordagem utilizada nesse trabalho. Planeja-se também buscar padrões para outros tamanhos de janelas de tempo, como por exemplo: janelas semanais, mensais e anuais. E, por fim, um melhor planejamento e análise de experimentos (RIBEIRO *et al.*, 2019). E, ainda, o uso da técnica de redução de dimencionalidades, a análise de componentes principais, para compressão dos dados.

6.2 Classificação - QP #2

Apresentamos também um método baseado em classificação para prever o nível de bem estar das colônias de abelhas que pode ser personalizado e integrado a um sistema de monitoramento via sensores de apiários (**Contribuição #2**). A partir dos dados dos sensores e dos dados climáticos externos, foi possível inferir com alta precisão, até **98**%, o nível de bem estar de uma colônia de abelhas (resposta para a **QP #2**). Além disso, uma das principais conclusões deste trabalho é que os dados do sensor e os dados de inspeção podem ser mesclados para alertar o apicultor sobre os problemas no bem estar das colônias. Portanto, o método aplicado pode ajudar decisivamente o apicultor a evitar perdas de colônias e auxiliar no manejo correto de suas colmeias. Também pode ajudar a impedir a diminuição das perdas de colônias durante o inverno, a estação do ano em que as inspeções geralmente não são viáveis e em que ocorre mais perdas. Assim, o método baseado em classificação também mitiga o **problema** apresentado na Seção 1.2 e confirma a **Hipótese #2** (Seção 1.3.1).

A partir do grande conjunto de dados obtido foi possível obter também informações úteis para os apicultores, tais como valores típicos de temperatura interna e peso da colmeia de acordo com as estações do ano e de acordo com os níveis de bem estar, bem como da temperatura externa das colmeias por apiário, estação do ano e o nível de bem estar. Essas informações são utilizadas pelos algoritmos de classificação para criar os modelos de classificação, contudo, são úteis também para o apicultor que dispõe de um sistema de monitoramento, mas não de um sistema de recomendação. Foi possível ainda extrair informações a partir das planilhas de inspeção, tais como: número de colônias doentes ao longos dos anos/estações e o percentual da ocorrência dos itens da planilha de inspeção, que é útil para indicar qual ou quais itens devem receber mais atenção pelo apicultor durante os manejos preventivos e/ou inspeções.

Em estudos futuros, planeja-se usar conjuntos de dados adicionais, como por exem-

plo, dados de colônias de abelhas africanizadas (*Apis mellifera*) no Brasil e de colônias em *Newfoundland*, Canadá. Outro possível estudo futuro envolve a criação de um sistema de sugestões de melhores práticas com base em ocorrências de estados não saudáveis. Essas práticas serão baseadas nas notas provenientes das inspeções dos apicultores. Essas anotações podem ser analisadas por meio da mineração de texto para procurar termos recorrentes para posterior classificação das melhores práticas. Também planejamos treinar, validar e testar um modelo do tipo *ensemble*, obtido a partir de modelos mesclados para criar um modelo mesclado, mais preciso e robusto. Pretende-se também realizar uma análise via regra de associação dos itens de inspeção. Poderia-se ainda usar um método baseado em série temporal para realização das remoção de anomalias. Também espera-se aplicar o método aqui apresentado em um conjunto de dados em que a atividade das abelhas rainhas é monitorada por sensores de áudio, tags RFID e sensores de umidade (ABOU-SHAARA *et al.*, 2017) e a presença de estressores também monitora através de sensores de gases (SZCZUREK *et al.*, 2019; SERITAN *et al.*, 2018).

6.3 Regressão - QP #3

Por fim, foi aplicado um método baseado em regressão para detectar a perda da capacidade de termorregulação de colônias de abelhas (**Contribuição #3**). Para o método baseado em regressão, os resultados obtidos mostraram que é possível prever quando as colônias de abelhas perdem sua capacidade de termorregular. O algoritmo utilizado foi eficiente em prever a perda de homeostase no ninho com algum tempo de antecedência. O algoritmo foi configurado para realizar a previsão em 2, 4, 10, 20, 24 e até 28 horas (janela de entrada) e, no caso das colmeias aqui estudadas, a solução proposta é capaz de prever com boa precisão a tendência de perda da capacidade de termorregulação de uma colônia com até 10 horas de antecedência com um erro de apenas **0.5**% (resposta para a **QP #3**). Portanto, um apicultor terá uma boa previsão do momento em que a temperatura começar a cair dentro do ninho, demonstrando que as abelhas não são mais capazes de termorregular o ninho. Vale ressaltar que o algoritmo utilizado realiza previsões em séries temporais. Assim, com o algoritmo treinado, ele pode ser usado em outras colmeias para prever estados eminentes de perda do controle termorregulatório. Assim, o método baseado em regressão também mitiga o **problema** apresentado na Seção 1.2 e a confirma a **Hipótese #3** definida na Seção 1.3.1.

Como perspectiva futura, pretende-se informar outros atributos ambientais oriundos de estações meteorológicas ao LSTM, tais como: temperatura do ponto de orvalho, direção do

vento, velocidade do vento, percipitação, luz do dia para obter os padrões de perda da capacidade de termorregulação e definir as possíveis razões para acontecer isso em um ambiente específico. Deseja-se também utilizar o nível de atividade da abelhas como atributo (GOMES *et al.*, 2020). Também planeja-se criar uma interface que permita notificar antecipadamente os apicultores quando uma colônia começar a perder a habilidade de termorregulação. Assim, o apicultor pode intervir e também determinar causas que levam à perda da termorregulação e aplicar as melhores práticas para ajudar a colônia.

6.4 Lista de Publicações

A seguir, são listados os principais trabalhos que foram desenvolvidos durante o curso de doutorado. Os trabalhos marcados com um "*"são os mais diretamente relacionados aos principais assuntos desta tese, mas todos foram relevantes no tocante à formação acadêmica.

(i) Artigo completo em periódicos

(1) * **BRAGA, ANTONIO RAFAEL**; GOMES, D. G.; ROGERS, R.; HASSLER, E. E.; FREITAS, B. M.; CAZIER, J. A. A Method for Mining Combined Data from in-Hive Sensors, Weather and Apiary Inspections to Forecast the Health Status of Honey Bee Colonies. Computers and Electronics in Agriculture, v. 169, p. 105161, 2020. ISSN 0168-1699.

(ii) Artigo completo em anais de conferências

(1) **BRAGA, ANTONIO RAFAEL**; MACIEL, F. A. O.; ALMEIDA, R. L. A.; AGUILAR, P. A. C.; GOMES, D.G.; ANDRADE, R. M. C. Gerenciamento Térmico e Elétrico de um Centro de Dados utilizando Sensoriamento IoT. In: Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP), 2017, São Paulo. Anais do 9º Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP), 2017. v. Único.

(2) MACIEL, F. A. O.; **BRAGA, ANTONIO RAFAEL**; SILVA, A. L. E.; SILVA, T. L. C.; FREITAS, B. M.; GOMES, D.G.. Reconhecimento de Padrões de Colônias de Abelhas Apis Mellifera Segundo Mudanças das Estações do Ano. In: CSBC 2018 - Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (9º WCAMA), 2018, Natal. XXXVIII Congresso da Sociedade Brasileira de Computação (CSBC), 2018.

(3) MACIEL, F. A. O.; **BRAGA, ANTONIO RAFAEL**; XAVIER, R. M.; DA SILVA, T. L. C.; FREITAS, B. M.; GOMES, D. G. Data Mining to Characterize Seasonal

Patterns of Apis mellifera Honey Bee Colonies. In: the XIV Brazilian Symposium, 2018, Caxias do Sul. Proceedings of the XIV Brazilian Symposium on Information Systems - SBSI'18. New York: ACM Press, 2018. p. 1.

(4) SILVA, W. F.; REGO, P. A. L.; MATEUS, B. G.; **BRAGA, ANTONIO RAFAEL**; ALENCAR, J. M. U. . DVL: Uma Ferramenta para Criação de Laboratórios Práticos de Disciplinas da Área de TI Utilizando Virtualização Baseada em Contêineres. In: SBRC 2018 - Salão de Ferramentas, 2018, Campus do Jordão. SBRC 2018 - Salão de Ferramentas, 2018.

(5) **BRAGA, ANTONIO RAFAEL**; FURTADO, L. S.; BEZERRA, A. D. M.; FREITAS, B. M.; CAZIER, J. A.; GOMES, D.G.. Applying the Long-Term Memory Algorithm to Forecast Loss of Thermoregulation Capacity in Honeybee Colonies. In: 10o Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais (WCAMA_CSBC 2019), 2019, Belém-PA. X Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais. Porto Alegre: Sociedade Brasileira de Computação, 2019. v. 10. p. 77-86.

(6) *** BRAGA, ANTONIO RAFAEL**; HASSLER, E. E.; GOMES, D.G.; FREITAS, B. M.; CAZIER, J. A.. IoT for Development: Building a Classification Algorithm to Help Beekeepers Detect Honeybee Health Problems Early. In: Twenty-fifth Americas Conference on Information Systems, 2019, Cancún. AMERICAS CONFERENCE ON INFORMATION SYSTEMS (AMCIS), 2019. p. 1-10.

(iii) Resumo expandido em anais de conferências

(1) *** BRAGA, ANTONIO RAFAEL**; SILVA, D. A.; NOBRE, J. S.; FREITAS, B. M.; GOMES, D.G.. Definindo e Predizendo Níveis de Saúde de Colônias de Abelhas via Clusterização e Classificação. In: Simpósio Brasileiro de Banco de Dados, 2019, Fortaleza. Simpósio Brasileiro de Banco de Dados, 2019.

(iv) Resumo em anais de conferências

(1) XAVIER, R. M.; **BRAGA, ANTONIO RAFAEL**; GOMES, D. G. . Classificação e Análise do Estado de Colônias de Abelhas Apis Mellifera. In: Encontros Universitários da UFC, 2018, Fortaleza. Encontros Universitários da UFC, 2018.

(2) WILLIAMS, J.; SCOTT, A.; ROGERS, R.; LINTON, F.; HASSLER, E.; **BRAGA, ANTONIO RAFAEL**; WILKES, J.; CAZIER, J. A.. Standardized Human Assessments of Hive Health: Assessing the Healthy Colony Checklist and Hooper?s 5

Questions as a Basis for a Standardized Bee Health Monitoring Tool. In: Apimondia International Apicultural Congress, 2019, Montréal. The 46th Apimondia International Apicultural Congress, 2019. v. 46. p. 237-237.

(3) SCOTT, A.; HASSLER, E.; **BRAGA, ANTONIO RAFAEL**; RUBINIGG, M. ; FORMATO, G.; WILKES, J.; CAZIER, J. A.. Software vs. Surveys: Comparing Approaches for Mapping Honey Bee Diseases. In: The 46th Apimondia International Apicultural Congress, 2019, Montréal. The 46th Apimondia International Apicultural Congress, 2019. v. 46. p. 238-238.

(4) * SILVA, D. A.; **BRAGA, ANTONIO RAFAEL**; NOBRE, J. S.; GOMES, D.G.. Clustering and Elastic Net Logistic Regression as Support Tools for Honeybee (*Apis mellifera*) Colonies Health Diagnosis. In: II Conference on Statistics and Data Science, 2019, Salvador. Conference on Statistics and Data Science, 2019.

6.5 Conclusões Gerais e Perspectivas

Embora os métodos aplicados nos capítulos anteriores tenham sido validados com apenas uma espécie de abelha (*Apis mellifera*), assumimos que estes métodos e os resultados obtidos são reproduzíveis e replicáveis desde que aplicados em condições semelhantes, ou seja, espécie da abelha, localização geográfica, estação do ano, dentre outros parâmetros utilizados durante a valiação dos métodos. Vale destacar ainda que, os médotos aplicados, são completamente aplicáveis para outras espécies de abelhas. Assim, em tese, é possível obter padrões sazonais, níveis de bem estar ou padrões termorregulatórios em colônias de outras regiões geográficas. Sugere-se, então, como pespectiva futura, que os 3 métodos apresentados sejam aplicados em outras espécies de abelhas, como por exemplo: as melíponas e as Bombus. Vale destacar ainda que as taxas de acerto obtidas são completamente aceitáveis para as predições que foram feitas.

O uso de sensores é, sem dúvida, benéfica para o bem estar as colônias, pois reduz o estresse causado pelas inspeções *in loco* e por possibilitar que apicultores possa saber com antecedência, como mostramos nessa tese, quando as colônias podem estar em iminente estado de colapso. Contudo, é importante reçaltar também como se dar em termos de tempo ou de dinheiro o retorno que o uso de sensores pode oferecer para um apicultor, expressa com que rapidez os investimentos para a implementação do sistema de medição específico serão retornados. Nessa linha, Zacepins *et al.* (2013) apresentaram um modelo matemático para determinar de rentabilidade da implementação de sistemas de monitoramento via sensores na agricultura de

precisão, ou seja, o conhecido Retorno Sobre o Investimento (ROI). Os autores fizeram essa estimativa em termos de tempo para esse retorno e observaram que para cada grandeza sensoriada é possível se obter um ROI específico. Os autores avaliaram quatro sensores, a saber: temperatura, som, vídeo e peso. Para os três primeiros sensores o tempo de retorno é de, respectivamente, 1.1, 1.6 e 1.5, anos. No caso do peso, o ROI pode ser de até 20 anos.

As mudanças climáticas e o uso de pesticidas podem impactar o bem estar das colônias e comprometer a manutenção de algumas espécies inclusive, assim, deseja-se ainda realizar estudo complementar de como tem sido nos últimos anos o impacto das mudanças climáticas, em especial, na produção agrícula relacionada à abelhas e à produção de produtos apícolas. Por fim, essa pesquisa ressalta a importância da ciência de dados bem como os tipos de metodologias propostas e validadas que podem ser integradas a uma operação de apicultura para ajudar as abelhas e os serviços de polinização que eles fornecem. Ao ajudar as colmeias a se tornarem melhor gerenciáveis através da fixação de sensores, é possível monitorá-las remotamente e com mais precisão ao longo do tempo. Assim, à medida que se aprende e interpreta os dados de sensores é possível usar as informações obtidas para rastrear mudanças de estado das colmeias e melhorar as operações dos apiários. Isso ajudará a evolução de colmeias inertes, inteligentes para, eventualmente, colmeias geniais (CAZIER, 2018).

REFERÊNCIAS

ABOU-SHAARA, H.; OWAYSS, A.; IBRAHIM, Y.; BASUNY, N. A review of impacts of temperature and relative humidity on various activities of honey bees. **Insectes Sociaux**, Springer, v. 64, n. 4, p. 455–463, 2017.

ABOU-SHAARA, H. F.; AL-GHAMDI, A. A.; MOHAMED, A. A. Tolerance of two honey bee races to various temperature and relative humidity gradients. **Environmental and experimental Biology**, v. 10, n. 4, p. 133–138, 2012.

AIZEN, M. A.; HARDER, L. D. The global stock of domesticated honey bees is growing slower than agricultural demand for pollination. **Current Biology**, v. 19, n. 11, p. 915 – 918, 2009. ISSN 0960-9822.

BECHER, M. A. **The influence of developmental temperatures on division of labour in honeybee colonies**. Tese (Doutorado) — Halle (Saale), Martin-Luther-Universität Halle-Wittenberg, 2010.

BEN-HUR, A.; HORN, D.; SIEGELMANN, H. T.; VAPNIK, V. Support vector clustering. **J. Mach. Learn. Res.**, JMLR.org, v. 2, p. 125–137, mar. 2002. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944790.944807>.

BENCSIK, M.; BENCSIK, J.; BAXTER, M.; LUCIAN, A.; ROMIEU, J.; MILLET, M. Identification of the honey bee swarming process by analysing the time course of hive vibrations. **Computers and Electronics in Agriculture**, v. 76, n. 1, p. 44 – 50, 2011. ISSN 0168-1699. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0168169911000068>.

BENCSIK, M.; CONTE, Y. L.; REYES, M.; PIOZ, M.; WHITTAKER, D.; CRAUSER, D.; DELSO, N. S.; NEWTON, M. I. Honeybee colony vibrational measurements to highlight the brood cycle. **PLOS ONE**, Public Library of Science, v. 10, n. 11, p. 1–16, 11 2015.

BEZERRA, A. D. M.; PACHECO FILHO, A. J. S.; BOMFIM, I. G. A.; SMAGGHE, G.; FREITAS, B. M. Agricultural area losses and pollinator mismatch due to climate changes endanger passion fruit production in the neotropics. **Agricultural Systems**, v. 169, n. C, p. 49–57, 2019.

BRAGA, A. R.; FURTADO, L.; BEZERRA, A. D.; FREITAS, B. M.; CAZIER, J. A.; GOMES, D. G. Applying the long-term memory algorithm to forecast thermoregulation capacity loss in honeybee colonies. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS, 10., 2019, Belém. **Anais do X Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais.** São Paulo: Sociedade Brasileira de Computação, 2019. p. 998–1007. Disponível em: <https://sol.sbc.org.br/index.php/wcama/article/download/6422/6318/>.

BRAGA, A. R.; GOMES, D. G.; ROGERS, R.; HASSLER, E. E.; FREITAS, B. M.; CAZIER, J. A. A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies. **Computers and Electronics in Agriculture**, v. 169, p. 105161, 2020. ISSN 0168-1699.

BRAGA, A. R.; HASSLER, E. E.; GOMES, D. G.; FREITAS, B. M.; CAZIER, J. A. Iot for development: Building a classification algorithm to help beekeepers detect honeybee health problems early. In: SPANISH, PORTUGUESE, AND LATIN AMERICA (LACAIS

CHAPTER), 6., 2019, Cancún. **Proceedings of Americas Conference on Information Systems.** Cancún: Association for Information Systems, 2019. p. 36:1–36:10. Disponível em: <https://aisel.aisnet.org/amcis2019/spanish_portuguese_latin_america/spanish_portuguese_latin_america/36>.

BRAGA, A. R.; MACIEL, F. A. O.; ALMEIDA, R. L. A.; AGUILAR, P. A. C.; GOMES, D. G.; ANDRADE, R. M. C. Gerenciamento térmico e elétrico de um centro de dados utilizando sensoriamento iot. In: SIMPOSIO BRASILEIRO DE COMPUTAÇÃO UBÍQUA E PERVASIVA, 9., São Paulo. **Anais do XXXVII Congresso da Sociedade Brasileira de Computação.** São Paulo: Sociedade Brasileira de Computação, 2017. p. 998–1007. Disponível em: <https://portaldeconteudo.sbc.org.br/index.php/sbcup/article/view/3306>.

BRAGA, A. R.; SILVA, D. A.; NOBRE, J. S.; FREITAS, B. M.; GOMES, D. G. Definindo e predizendo níveis de saúde de colônias de abelhas via clusterização e classificação. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 34., 2019, Fortaleza. **Anais do XXXIV Simpósio Brasileiro de Banco de Dados.** Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 241–246. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/8830>.

BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R. T.; SANDER, J. Lof: Identifying density-based local outliers. In: ACM SIGMOD CONFERENCE, 26., 2000, Dalles. **Proceedings of the International Conference on Management of Data.** New York: Association for Computing Machinery, 2000. p. 93–104. ISBN 1-58113-217-4. Disponível em: <http://doi.acm.org/10.1145/342009.335388>.

BRODSCHNEIDER, R.; GRAY, A.; ADJLANE, N.; BALLIS, A.; BRUSBARDIS, V.; CHARRIèRE, J.-D.; CHLEBO, R.; COFFEY, M. F.; DAHLE, B.; GRAAF, D. C. de; DRAžIć, M. M.; EVANS, G.; FEDORIAK, M.; FORSYTHE, I.; GREGORC, A.; GRZEDA, U.; HETZRONI, A.; KAUKO, L.; KRISTIANSEN, P.; MARTIKKALA, M.; MARTíN-HERNáNDEZ, R.; MEDINA-FLORES, C. A.; MUTINELLI, F.; RAUDMETS, A.; RYZHIKOV, V. A.; SIMON-DELSO, N.; STEVANOVIC, J.; UZUNOV, A.; VEJSNæS, F.; WöHL, S.; ZAMMIT-MANGION, M.; DANIHLíK, J. Multi-country loss rates of honey bee colonies during winter 2016/2017 from the coloss survey. **Journal of Apicultural Research**, Taylor & Francis, v. 57, n. 3, p. 452–457, 2018.

CALDERONE, N. W. Insect pollinated crops, insect pollinators and us agriculture: trend analysis of aggregate data for the period 1992–2009. **PloS one**, Public Library of Science, v. 7, n. 5, 2012.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.

CARVALHO, H. V. F. de; CARVALHO, E. C.; ARRUDA, H.; IMPERATRIZ-FONSECA, V.; SOUZA, P. de; PESSIN, G. Detecção de anomalias em comportamento de abelhas utilizando redes neurais recorrentes. In: WORKSHOP DE COMPUTAÇÃO APLICADA A GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS, 9., 2018, Natal. **Anais do XXXVIII Congresso da Sociedade Brasileira de Computação.** Porto Alegre: Sociedade Brasileira de Computação, 2018. p. 77–86. ISSN 2595-6124. Disponível em: <http://portaldeconteudo.sbc.org.br/index.php/wcama/article/view/2931>.

CAVALCANTE, M. C.; GALETTO, L.; MAUÉS, M. M.; PACHECO FILHO, A. J. S.; BOMFIM, I. G. A.; FREITAS, B. M. Nectar production dynamics and daily pattern of pollinator visits in brazil nut (bertholletia excelsa bonpl.) plantations in central amazon: implications for fruit production. **Apidologie**, Jun 2018. ISSN 1297-9678. Disponível em: <https://doi.org/10.1007/s13592-018-0578-y>.

CAZIER, J. A. **Peering Into The Future A Path To The Genius Hive | Bee Culture**. 2018. 44 - 46 p. Pages 44-46. Accessed 14 Dec. 2018. Disponível em: <https://www.beeculture.com/peering-into-the-future-a-path-to-the-genius-hive/>.

CHAZETTE, L.; BECKER, M.; SZCZERBICKA, H. Basic algorithms for bee hive monitoring and laser-based mite control. In: IEEE SYMPOSIUM SERIES ON COMPUTATIONAL INTELLIGENCE (SSCI), 6., 2016, Athens. **Proceedings of Symposium Series on Computational Intelligence.** Piscataway: Institute of Electrical and Electronics Engineers, 2016. p. 1–8. Disponível em: <https://ieeexplore.ieee.org/document/7850001>.

CHEN, C.; YANG, E.-C.; JIANG, J.-A.; LIN, T.-T. An imaging system for monitoring the in-and-out activity of honey bees. **Computers and Electronics in Agriculture**, v. 89, p. 100 – 109, 2012. ISSN 0168-1699. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0168169912002074>.

CHEN, M.-S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and data Engineering**, IEEE, v. 8, n. 6, p. 866–883, 1996.

COOK, C. N.; BREED, M. D. Social context influences the initiation and threshold of thermoregulatory behaviour in honeybees. **Animal Behaviour**, Elsevier, v. 86, n. 2, p. 323 – 329, 2013. ISSN 0003-3472.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, January 1967. ISSN 1557-9654.

CRANE, E. Recent research on the world history of beekeeping. **Bee World**, Taylor & Francis, v. 80, n. 4, p. 174–186, 1999.

CURRIE, R. W.; PERNAL, S. F.; GUZMáN-NOVOA, E. Honey bee colony losses in canada. **Journal of Apicultural Research**, Taylor & Francis, v. 49, n. 1, p. 104–106, 2010.

DINEVA, K.; ATANASOVA, T. Applying machine learning against beehives dataset. **International Multidisciplinary Scientific GeoConference: SGEM: Surveying Geology & mining Ecology Management**, Surveying Geology & Mining Ecology Management (SGEM), v. 18, p. 35–42, 2018.

DINEVA, K.; ATANASOVA, T. Osemn process for working over data acquired by iot devices mounted in beehives. **Current Trends in Natural Sciences**, v. 7, p. 47–53, 01 2018.

EFSA Panel on Animal Health and Welfare (AHAW). Assessing the health status of managed honeybee colonies (healthy-b): a toolbox to facilitate harmonised data collection. **EFSA Journal**, v. 14, n. 10, p. e04578, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FISHER, R. A. The correlation between relatives on the supposition of mendelian inheritance. **Transactions of the Royal Society of Edinburgh**, Royal Society of Edinburgh Scotland Foundation, v. 52, n. 2, p. 399–433, 1918.

FITZGERALD, D. W.; MURPHY, F. E.; WRIGHT, W. M. D.; WHELAN, P. M.; POPOVICI, E. M. Design and development of a smart weighing scale for beehive monitoring. In: IRISH SIGNALS AND SYSTEMS CONFERENCE (ISSC), 26., 2015, Ireland. **Proceedings of Irish Signals and Systems Conference (ISSC).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2015. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/7163763>.

FLORES, J. M.; GIL-LEBRERO, S.; GáMIZ, V.; RODRíGUEZ, M. I.; ORTIZ, M. A.; QUILES, F. J. Effect of the climate change on honey bee colonies in a temperate mediterranean zone assessed through remote hive weight monitoring system in conjunction with exhaustive colonies assessment. **Science of The Total Environment**, v. 653, p. 1111 – 1119, 2019. ISSN 0048-9697.

FORMATO, G.; SMULDERS, F. J. Risk management in primary apicultural production. part 1: bee health and disease prevention and associated best practices. **Veterinary Quarterly**, Taylor & Francis, v. 31, n. 1, p. 29–47, 2011. PMID: 22029819.

FREITAS, B. M.; SOUSA, R.; BOMFIM, I. Absconding and migratory behaviors of feral africanized honey bee (apis mellifera l.) colonies in ne brazil = comportamentos de abandono e migração de colônias silvestres da abelha melífera africanizada (apis mellifera l.) no nordeste do brasil. **Acta Scientiarum : Biological Sciences**, v. 29, 10 2007.

GALLAI, N.; SALLES, J.-M.; SETTELE, J.; VAISSIèRE, B. E. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. **Ecological Economics**, v. 68, n. 3, p. 810 – 821, 2009. ISSN 0921-8009.

GARIBALDI, L. A. *et al.* Wild pollinators enhance fruit set of crops regardless of honey bee abundance. **Science (New York, N.Y.)**, The American Association for the Advancement of Science, v. 339, n. 6127, p. 1608–1611, 2013. ISSN 1095-9203.

GIANNINI, T. C.; CORDEIRO, G. D.; FREITAS, B. M.; SARAIVA, A. M.; IMPERATRIZ-FONSECA, V. L. The dependence of crops for pollinators and the economic value of pollination in brazil. **Journal of Economic Entomology**, Oxford University Press, v. 108, n. 3, p. 849–857, 2015.

GIL-LEBRERO, S.; QUILES-LATORRE, F. J.; ORTIZ-LóPEZ, M.; SáNCHEZ-RUIZ, V.; GáMIZ-LóPEZ, V.; LUNA-RODRíGUEZ, J. J. Honey bee colonies remote monitoring system. **Sensors**, v. 17, n. 1, 2017. ISSN 1424-8220. Disponível em: <http://www.mdpi.com/1424-8220/17/1/55>.

GILIOLI, G.; SPERANDIO, G.; HATJINA, F.; SIMONETTO, A. Towards the development of an index for the holistic assessment of the health status of a honey bee colony. **Ecological Indicators**, v. 101, p. 341 – 347, 2019. ISSN 1470-160X.

GOMES, P.; SUHARA, Y.; NUNES-SILVA, P.; COSTA, L.; ARRUDA, H.; VENTURIERI, G.; IMPERATRIZ-FONSECA, V.; PENTLAND, A.; SOUZA, P. D.; PESSIN, G. An amazon stingless bee foraging activity predicted using recurrent artificial neural networks and attribute selection. **Scientific Reports**, v. 10, p. 9, 12 2020.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, Sep. 2009. ISSN 1041-4347.

HELDMAIER, G. Temperature control in honey bee colonies. **BioScience**, v. 37, p. 395–399, 06 1987.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. **Neural Comput.**, MIT Press, Cambridge, MA, USA, v. 18, n. 7, p. 1527–1554, jul. 2006. ISSN 0899-7667.

HO, T. K. Random decision forests. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 3., 1995, Montreal. **Proceedings of 3rd International Conference on Document Analysis and Recognition.** Piscataway: Institute of Electrical and Electronics Engineers, 1995. p. 278–282. Disponível em: <http://doi.org/10.1109/ICDAR.1995.598994>.

HOOPER, T. Book. **Guide to bees and honey**. London, UK: Blandford Poole, Dorset, 1996. 260 p. p. ISBN 0713707828.

HU, M. Y.; ZHANG, G. P.; JIANG, C. X.; PATUWO, B. E. A cross-validation analysis of neural network out-of-sample performance in exchange rate forecasting. **Decision Sciences**, v. 30, n. 1, p. 197–216, 1999.

JACOBS, M.; CAZIER, J. A.; WILKES, J. T.; ROGERS, R.; HASSLER, E. E. Building a business analytics platform for enhancing commercial beekeepers' performance: Descriptive validation of a data framework for widespread adoption by citizen scientists. In: AMERICAS CONFERENCE ON INFORMATION SYSTEMS, 23., 2017, Boston. **Proceedings of 23rd Americas Conference on Information Systems.** Boston: Association for Information Systems, 2017. p. 1–10. Disponível em: <http://aisel.aisnet.org/amcis2017/DataScience/Presentations/24>.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, Inc., 1988.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: A review. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, v. 22, n. 1, p. 4–37, 2000.

KEVAN, P.; PHILLIPS, T. The economic impacts of pollinator declines: an approach to assessing the consequences. **Conservation Ecology**, The Resilience Alliance, v. 5, n. 1, 2001.

KLEIN, A.-M.; VAISSIERE, B. E.; CANE, J. H.; STEFFAN-DEWENTER, I.; CUNNINGHAM, S. A.; KREMEN, C.; TSCHARNTKE, T. Importance of pollinators in changing landscapes for world crops. **Proceedings of the Royal Society of London B: Biological Sciences**, The Royal Society, v. 274, n. 1608, p. 303–313, 2007.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995, Montreal. [S.l.]: **Proceedings of the 14th international joint conference on Artificial intelligence.** San Francisco: Morgan Kaufmann, 1995. p. 1137–1143.

KREYSZIG, E. **Advanced Engineering Mathematics**. [S.l.]: John Wiley & Sons, 2010. 1036–1038 p. ISBN 9780470458365.

KRIDI, D. S.; CARVALHO, C. G. N.; GOMES, D. G. A predictive algorithm for mitigate swarming bees through proactive monitoring via wireless sensor networks. In: SYMPOSIUM ON PERFORMANCE EVALUATION OF WIRELESS AD HOC, SENSOR, AND UBIQUITOUS NETWORKS, 11., 2014, Montreal. **Proceedings of the 11th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks.** New York: Association for Computing Machinery (ACM), 2014. p. 41–47. ISBN 978-1-4503-3025-1. Disponível em: <http://doi.acm.org/10.1145/2653481.2653482>.

KRIDI, D. S.; CARVALHO, C. G. N. de; GOMES, D. G. Application of wireless sensor networks for beehive monitoring and in-hive thermal patterns detection. **Computers and Electronics in Agriculture**, v. 127, p. 221 – 235, 2016. ISSN 0168-1699.

KULHANEK, K.; STEINHAUER, N.; RENNICH, K.; CARON, D. M.; SAGILI, R. R.; PETTIS, J. S.; ELLIS, J. D.; WILSON, M. E.; WILKES, J. T.; TARPY, D. R.; ROSE, R.; LEE, K.; RANGEL, J.; VANENGELSDORP, D. A national survey of managed honey bee 2015–2016 annual colony losses in the usa. **Journal of Apicultural Research**, Taylor & Francis, v. 56, n. 4, p. 328–340, 2017.

KVIESIS, A.; KOMASILOVS, V.; KOMASILOVA, O.; ZACEPINS, A. Application of fuzzy logic for honey bee colony state detection based on temperature data. **Biosystems Engineering**, v. 193, p. 90 – 100, 2020. ISSN 1537-5110. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1537511020300507>.

KVIESIS, A.; ZACEPINS, A. Application of neural networks for honey bee colony state identification. In: CARPATHIAN CONTROL CONFERENCE (ICCC), 17., 2016, High Tatras. [S.l.]: **Proceeding of 17th International Carpathian Control Conference (ICCC).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2016. p. 413–417.

LEE, K.; STEINHAUER, N.; TRAVIS, D. A.; MEIXNER, M. D.; DEEN, J.; VANENGELSDORP, D. Honey bee surveillance: a tool for understanding and improving honey bee health. **Current Opinion in Insect Science**, v. 10, p. 37 – 44, 2015. ISSN 2214-5745. Social Insects * Vectors and Medical and Veterinary Entomology.

LEVÉ, M.; BAUDRY, E.; BESSA-GOMES, C. Domestic gardens as favorable pollinator habitats in impervious landscapes. **Science of The Total Environment**, v. 647, p. 420 – 430, 2019. ISSN 0048-9697.

MACIEL, F. A.; BRAGA, A. R.; SILVA, T. L. C. da; FREITAS, B. M.; GOMES, D. Reconhecimento de padrões sazonais em colônias de abelhas apis mellifera via clusterização. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 74–88, out. 2018. Disponível em: <http://seer.upf.br/index.php/rbca/article/view/8788>.

MACIEL, F. A. O.; BRAGA, A. R.; SILVA, A. L.; SILVA, T. L. C. da; FREITAS, B. M.; GOMES, D. G. Reconhecimento de padrões de colônias de abelhas apis mellifera segundo

mudanças das estações do ano. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 9. , 2018, Natal. **Anais do IX Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais.** Porto Alegre: Sociedade Brasileira de Computação (SBC), 2018. p. 1–10. ISSN 2595-6124. Disponível em: <https://sol.sbc.org.br/index.php/wcama/article/view/2937>.

MACIEL, F. A. O.; BRAGA, A. R.; XAVIER, R. M.; SILVA, T. L. C. da; FREITAS, B. M.; GOMES, D. G. Data mining to characterize seasonal patterns of apis mellifera honey bee colonies. In: BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 14., 2018, Caxias do Sul. **Proceedings of the XIV Brazilian Symposium on Information Systems.**, New York: Association for Computing Machinery (ACM), 2018. p. 38:1–38:8. ISBN 978-1-4503-6559-8. Disponível em: <http://doi-acm-org.ez11.periodicos.capes.gov.br/10.1145/3229345.3229386>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Berkeley. [S.l.]: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.** Berkeley: Univ. of Calif. Press, 1967. p. 281–297.

MALAGODI-BRAGA, K. S. A polinização como fator de produção na cultura do morango. **Embrapa Meio Ambiente-Comunicado Técnico (INFOTECA-E)**, Jaguariúna: Embrapa Meio Ambiente, 2018.

MARON, M. E. Automatic indexing: An experimental inquiry. **J. ACM**, Association for Computing Machinery, New York, NY, USA, v. 8, n. 3, p. 404–417, jul. 1961. ISSN 0004-5411.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, n. 4, p. 115–133, Dec 1943. ISSN 1522-9602. Disponível em: <https://doi.org/10.1007/BF02478259>.

MCNALLY, L. C.; SCHNEIDER, S. S. Seasonal cycles of growth, development and movement of the african honey bee, apis mellifera scutettata, in africa. **Insectes Sociaux**, v. 39, n. 2, p. 167–179, Jun 1992. ISSN 1420-9098. Disponível em: <https://doi.org/10.1007/BF01249292>.

MEIKLE, W. G.; HOLST, N. Application of continuous monitoring of honeybee colonies. **Apidologie**, v. 46, n. 1, p. 10–22, Jan 2015. ISSN 1297-9678.

MEIKLE, W. G.; WEISS, M.; MAES, P. W.; FITZ, W.; SNYDER, L. A.; SHEEHAN, T.; MOTT, B. M.; ANDERSON, K. E. Internal hive temperature as a means of monitoring honey bee colony health in a migratory beekeeping operation before and during winter. **Apidologie**, v. 48, n. 5, p. 666–680, Sep 2017. Disponível em: <https://doi.org/10.1007/s13592-017-0512-8>.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Springer, v. 50, n. 2, p. 159–179, 1985.

MORSE, R. A.; CALDERONE, N. W. The value of honey bees as pollinators of us crops in 2000. **Bee culture**, Medina, OH, v. 128, n. 3, p. 1–15, 2000.

MUMBI, C. T.; MWAKATOBE, A. R.; MPINGA, I. H.; RICHARD, A.; MACHUMU, R. Parasitic mite, varroa species (parasitiformes: Varroidae) infesting the colonies of african honeybees, apis mellifera scutellata (hymenoptera: Apididae) in tanzania. **J. Entomol. Zool. Stud**, v. 2, n. 3, p. 188–196, 2014.

MURPHY, F. E.; MAGNO, M.; O'LEARY, L.; TROY, K.; WHELAN, P.; POPOVICI, E. M. Big brother for bees (3b) — energy neutral platform for remote monitoring of beehive imagery and sound. In: IEEE INTERNATIONAL WORKSHOP ON ADVANCES IN SENSORS AND INTERFACES (IWASI), 6., 2015, Gallipoli. **Proceedings of the 6th IEEE International Workshop on Advances in Sensors and Interfaces (IWASI).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2015. p. 106–111. Disponível em: <https://ieeexplore.ieee.org/document/7184943>.

MURPHY, F. E.; MAGNO, M.; WHELAN, P. M.; O'HALLORAN, J.; POPOVICI, E. M. b+wsn: Smart beehive with preliminary decision tree analysis for agriculture and honey bee health monitoring. **Computers and Electronics in Agriculture**, v. 124, p. 211 – 219, 2016. ISSN 0168-1699.

MURPHY, F. E.; SRBINOVSKI, B.; MAGNO, M.; POPOVICI, E. M.; WHELAN, P. M. An automatic, wireless audio recording node for analysis of beehives. In: IRISH SIGNALS AND SYSTEMS CONFERENCE (ISSC), 26., 2015, Carlow. **Proceedings of the 26th Irish Signals and Systems Conference (ISSC).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2015. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/7163753>.

NEUPANE, K.; THAPA, R. Pollen collection and brood production by honeybees (apis mellifera l.) under chitwan condition of nepal. **Journal of the Institute of Agriculture and Animal Science**, v. 26, n. 0, p. 143–148, 2005.

OLLERTON, J. Pollinator diversity: Distribution, ecological function, and conservation. **Annual Review of Ecology, Evolution, and Systematics**, v. 48, n. 1, p. 353–376, 2017.

OLLERTON, J.; WINFREE, R.; TARRANT, S. How many flowering plants are pollinated by animals? **Oikos**, v. 120, n. 3, p. 321–326, 2011.

OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION (MLDM), 8., 2012, Berlin. [S.l.]: **Proceedings of the 8th International MLDM.** Berlin: Heidelberg Springer, 2012. p. 154–168. ISBN 978-3-642-31537-4.

OSTWALD, M. M.; SMITH, M. L.; SEELEY, T. D. The behavioral regulation of thirst, water collection and water storage in honey bee colonies. **Journal of Experimental Biology**, The Company of Biologists Ltd, v. 219, n. 14, p. 2156–2165, 2016.

PIATESKI, G.; FRAWLEY, W. **Knowledge Discovery in Databases**. Cambridge, MA, USA: MIT Press, 1991. ISBN 0262660709.

PIATETSKY-SHAPIRO, G. Knowledge discovery in real databases: A report on the ijcai-89 workshop. **AI Magazine**, v. 11, n. 4, p. 68, Dec. 1990.

POTTS, S.; IMPERATRIZ-FONSECA, V. L.; NGO, H. T.; BIESMEIJER, T. D. B. J. C.; DICKS, L. V.; GARIBALDI, L. A.; HILL, R.; SETTELE, J.; VANBERGEN, A. J.; AIZEN, M. A.; CUNNINGHAM, S. A.; EARDLEY, C.; FREITAS, B. M.; GALLAI, N.; KEVAN, P. G.; KOVáCS-HOSTYáNSZKI, A.; KWAPONG, P. K.; LI, J.; LI, X.; MARTINS, D. J.; NATES-PARRA, G.; PETTIS, J. S.; RADER, R.; VIANA, B. F. **Summary for policymakers of the assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production**. 1. ed. Bonn, Germany: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2016.

POTTS, S. G.; BIESMEIJER, J. C.; KREMEN, C.; NEUMANN, P.; SCHWEIGER, O.; KUNIN, W. E. Global pollinator declines: trends, impacts and drivers. **Trends in Ecology & Evolution**, Elsevier, v. 25, n. 6, p. 345 – 353, 2010. ISSN 0169-5347.

POTTS, S. G.; IMPERATRIZ-FONSECA, V.; NGO, H. T.; AIZEN, M. A.; BIESMEIJER, J. C.; BREEZE, T. D.; DICKS, L. V.; GARIBALDI, L. A.; HILL, R.; SETTELE, J.; VANBERGEN, A. J. Safeguarding pollinators and their values to human well-being. **Nature**, v. 540, p. 220–229, 2016.

RIBEIRO, J. E. E.; ZEVIANI, W. M.; BONAT, W. H.; DEMETRIO, C. G.; HINDE, J. Reparametrization of com–poisson regression models with applications in the analysis of experimental data. **Statistical Modelling**, v. 0, n. 0, p. 1471082X19838651, 2019.

RICE, L. **Wireless Data Acquisition for Apiology Applications**. Dissertação (Mestrado) — Appalachian State University, Boone, NC., 2013. Disponível em: <https://libres.uncg.edu/ir/asu/f/Rice,\%20Luke_2013_Thesis.pdf>.

ROBLES-GUERRERO, A.; SAUCEDO-ANAYA, T.; GONZáLEZ-RAMíREZ, E.; ROSA-VARGAS, J. I. D. la. Analysis of a multiclass classification problem by lasso logistic regression and singular value decomposition to identify sound patterns in queenless bee colonies. **Computers and Electronics in Agriculture**, v. 159, p. 69 – 74, 2019. ISSN 0168-1699.

ROGERS, R. **Healthy Colony Checklist (HCC) - Bayer Crop Science**. 2017. Https://beehealth.bayer.us/who-can-help/beekeepers/healthy-colony-checklist. Accessed on 07-March-2019.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958. ISSN 0033-295X.

RUAN, Z.-Y.; WANG, C.-H.; LIN, H.-J.; HUANG, C.-P.; CHEN, Y.-H.; YANG, E.-C.; TSENG, C.-L.; JIANG, J.-A. An internet of things-based weight monitoring system for honey. **International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering**, World Academy of Science, Engineering and Technology, v. 11, n. 6, p. 478 – 482, 2017. ISSN eISSN:1307-6892.

SAK, H.; SENIOR, A. W.; BEAUFAYS, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 15., 2014, Singapore. [S.l.]: **Proceedings of the 15th Annual Conference of the International Speech Communication Association.** Singapore: ISCA Archive, 2014. p. 338–342.

SÁNCHEZ, V.; GIL, S.; FLORES, J. M.; QUILES, F. J.; ORTIZ, M. A.; LUNA, J. J. Implementation of an electronic system to monitor the thermoregulatory capacity of honeybee colonies in hives with open-screened bottom boards. **Computers and Electronics in Agriculture**, v. 119, p. 209 – 216, 2015. ISSN 0168-1699.

SEELEY, T. D.; VISSCHER, P. K. Survival of honeybees in cold climates: the critical timing of colony growth and reproduction. **Ecological Entomology**, v. 10, n. 1, p. 81–88, 1985.

SERITAN, G. C.; ENACHE, B.; ARGATAU, F. C.; ADOCHIEI, F. C.; TOADER, S. Low cost platform for monitoring honey production and bees health. In: IEEE INTERNATIONAL

CONFERENCE ON AUTOMATION, QUALITY AND TESTING, ROBOTICS (AQTR), 9., 2018, Cluj-Napoca. [S.l.]: **Proceedings of 9th IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2018. p. 1–4.

SILVA, T. L. C. da; MACêDO, J. A. F. de; CASANOVA, M. A. Discovering frequent mobility patterns on moving object data. In: SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 30., 2014, Dallas. [S.l.]: **Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems.** New York: Association for Computing Machinery (ACM), 2014. p. 60–67. ISBN 978-1-4503-3142-5.

SPILIOPOULOU, M.; NTOUTSI, I.; THEODORIDIS, Y.; SCHULT, R. Monic: Modeling and monitoring cluster transitions. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 12., 2006, Philadelphia. [S.l.]: **Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.** New York: Association for Computing Machinery (ACM), 2006. p. 706–711. ISBN 1-59593-339-5.

SPIVAK, M.; REUTER, G. S. Resistance to american foulbrood disease by honey bee colonies apis mellifera bred for hygienic behavior. **Apidologie**, EDP Sciences, v. 32, n. 6, p. 555–565, 2001.

SPONSLER, D. B.; GROZINGER, C. M.; HITAJ, C.; RUNDLöF, M.; BOTíAS, C.; CODE, A.; LONSDORF, E. V.; MELATHOPOULOS, A. P.; SMITH, D. J.; SURYANARAYANAN, S.; THOGMARTIN, W. E.; WILLIAMS, N. M.; ZHANG, M.; DOUGLAS, M. R. Pesticides and pollinators: A socioecological synthesis. **Science of The Total Environment**, v. 662, p. 1012 – 1027, 2019. ISSN 0048-9697.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, v. 15, p. 1929–1958, 2014. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>.

STALIDZANS, E.; BERZONIS, A. Temperature changes above the upper hive body reveal the annual development periods of honey bee colonies. **Computers and Electronics in Agriculture**, v. 90, p. 1 – 6, 2013. ISSN 0168-1699.

STALIDZANS, E.; BILINSKIS, V.; BERZONIS, A. Determination of development periods of honeybee colony by temperature in hive in latvia, year 2000. **Apiacta**, 2002. Disponível em: <http://www.fiitea.org/foundation/files/2002/E.\%20STALIDZANS.pdf>.

SZCZUREK, A.; MACIEJEWSKA, M.; BAK, B.; WILDE, J.; SIUDA, M. Semiconductor gas sensor as a detector of varroa destructor infestation of honey bee colonies – statistical evaluation. **Computers and Electronics in Agriculture**, v. 162, p. 405 – 411, 2019. ISSN 0168-1699.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining, (First Edition)**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

TASHAKKORI, R.; GHADIRI, A. Image processing for honey bee hive health monitoring. In: IEEE SOUTHEASTCON, 30., 2015, Fort Lauderdale. **Proceedings of the 30th IEEE Southeastcon.** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2015. p. 1–7. ISSN 1558-058X. Disponível em: <https://ieeexplore.ieee.org/document/7133029>.

TUKEY, J. W. **Exploratory Data Analysis**. [S.l.]: Addison-Wesley Publishing Company, 1977. v. 2.

VANENGELSDORP, D.; MEIXNER, M. D. A historical review of managed honey bee populations in europe and the united states and the factors that may affect them. **Journal of Invertebrate Pathology**, v. 103, p. S80 – S95, 2010. ISSN 0022-2011.

VANENGELSDORP, D.; TARPY, D. R.; LENGERICH, E. J.; PETTIS, J. S. Idiopathic brood disease syndrome and queen events as precursors of colony mortality in migratory beekeeping operations in the eastern united states. **Preventive Veterinary Medicine**, v. 108, n. 2, p. 225 – 233, 2013. ISSN 0167-5877.

VERBOVEN, H. A.; UYTTENBROECK, R.; BRYS, R.; HERMY, M. Different responses of bees and hoverflies to land use in an urban–rural gradient show the importance of the nature of the rural land use. **Landscape and Urban Planning**, v. 126, p. 31 – 41, 2014. ISSN 0169-2046.

WALTON, E.; JACOBS, M.; WILKES, J. T.; CAZIER, J. A. Building a business analytics platform for enhancing commercial beekeepers' performance. In: AMERICAS CONFERENCE ON INFORMATION SYSTEMS (AMCIS), 22., 2016, San Diego. **Proceedings of the 22nd Americas Conference on Information Systems (AMCIS).** San Diego: Americas Information System (AIS), 2016. p. 1–10. Disponível em: <http://aisel.aisnet.org/amcis2016/Decision/Presentations/20>.

WINSTON, M. **The Biology of the Honey Bee**. Harvard University Press, 1991. ISBN 9780674074095. Disponível em: <https://books.google.com/books?id=-5iobWHLtAQC>.

WINSTON, M. L. Seasonal patterns of brood rearing and worker longevity in colonies of the africanized honey bee (hymenoptera: Apidae) in south america. **Journal of the Kansas Entomological Society**, Kansas (Central States) Entomological Society, v. 53, n. 1, p. 157–165, 1980. ISSN 00228567, 19372353. Disponível em: <http://www.jstor.org/stable/25084014>.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; PHILIP, S. Y. *et al.* Top 10 algorithms in data mining. **Knowledge and information systems**, Springer, v. 14, n. 1, p. 1–37, 2008.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. **Annals of Data Science**, v. 2, n. 2, p. 165–193, Jun 2015. ISSN 2198-5812. Disponível em: <https://doi.org/10.1007/s40745-015-0040-1>.

XU, R.; WUNSCH, D. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, Ieee, v. 16, n. 3, p. 645–678, 2005.

ZACEPINS, A. Application of bee hive temperature measurements for recognition of bee colony state. In: INTERNATIONAL SCIENTIFIC CONFERENCE APPLIED INFORMATION AND COMMUNICATION TECHNOLOGIES, 5., 2012, Jelgava. **Proceedings of the 5th International Scientific Conference Applied Information and Communication Technologies.** Jelgava: Latvia Univ. of Agriculture Press, 2012. p. 216–221. Disponível em: <http://agris.fao.org/agris-search/search.do?recordID=LV2012000422>.

ZACEPINS, A.; BRUSBARDIS, V.; MEITALOVS, J.; STALIDZANS, E. Challenges in the development of precision beekeeping. **Biosystems Engineering**, v. 130, p. 60 – 71, 2015. ISSN 1537-5110.

ZACEPINS, A.; KARASHA, T. Web based system for the bee colony remote monitoring. In: INTERNATIONAL CONFERENCE ON APPLICATION OF INFORMATION AND COMMUNICATION TECHNOLOGIES (AICT), 6., 2012, Georgia. **Proceedings of the 6th International Conference on Application of Information and Communication Technologies (AICT).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2012. p. 1–4. Disponível em: <https://ieeexplore.ieee.org/document/6398490>.

ZACEPINS, A.; KVIESIS, A.; PECKA, A.; OSADCUKS, V. Development of internet of things concept for precision beekeeping. In: INTERNATIONAL CARPATHIAN CONTROL CONFERENCE (ICCC), 18., 2017, Sinaia. [S.l.]: **Proceedings of the 18th International Carpathian Control Conference (ICCC).** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2017. p. 23–27.

ZACEPINS, A.; KVIESIS, A.; STALIDZANS, E.; LIEPNIECE, M.; MEITALOVS, J. Remote detection of the swarming of honey bee colonies by single-point temperature monitoring. **Biosystems Engineering**, v. 148, p. 76 – 80, 2016. ISSN 1537-5110. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1537511016300964>.

ZACEPINS, A.; STALIDZANS, E.; KARASHA, T. Profitability ranking of precision agriculture measurement system implementation. In: INTERNATIONAL SCIENTIFIC CONFERENCE ENGINEERING FOR RURAL DEVELOPMENT, 12., 2013, Jelgava. [S.l.]: **Proceedings of the 12th International Scientific Conference Engineering for Rural Development.** Latvia: University of Life Sciences and Technologies Press, 2013. p. 164–169.

ZOGOVIC, N.; MLADENOVIC, M.; RASIC, S. From primitive to cyber-physical beekeeping. In: INTERNATIONAL CONFERENCE ON INFORMATION SOCIETY AND TECHNOLOGY (ICIST), 7., 2017, Kopaonik. [S.l.]: **Proceedings 7th International Conference on Information Society and Technology.** Belgrade: Society for Information Systems and Computer Networks, 2017. p. 38–43.

ŽGANK, A. Acoustic monitoring and classification of bee swarm activity using mfcc feature extraction and hmm acoustic modeling. In: IEEE-ELEKTRO CONFERENCE, 15., 2018, Mikulov. [S.l.]: **IEEE-Elektro Conference.** Piscataway: Institute of Electrical and Electronics Engineers (IEEE), 2018. p. 1–4.

ANEXO A – LISTAS DE ITENS DE VERIFICAÇÃO DA SAÚDE DAS COLÔNIAS

Bayer Bee Care Center: Bee Health Integrated Apiculture Research Program

Science For A Better Life

Page _____ of _____

HEALTHY COLONY CHECKLIST

This checklist is useful for quick assessments anytime a hive is opened, and as a summary of more detailed assessments. The results should answer the questions 1) Is the colony healthy?, 2) If not, why?, and 3) What needs to be done before the next weekly assesment to correct the problem?

Date:					
Apiary ID:		**Hive ID:**		**Number frames in brood box 2 (upper):**	
Observer:		**Recorder:**		**Number of frames in brood box 1 (lower):**	

For a colony to be considered "healthy", it must satisfy **ALL** of the following conditions, as seasonally appropriate.

Condition Met? *	Condition to Assess	Notes, Problems Observed & Management Needed
	1 - All stages of brood and instars present in appropriate amounts (Eggs 1-3, Larvae 1-6, Pupae 1-11)?	
	2 - Sufficient adult bees and age structure to care for brood and perform all tasks of the colony?**	
	3 - A young (<1 yr old), productive, laying **queen present**? *(Color Code Guide: Blue(0/5), White(1/6), Yellow(2/7), Red(3/8), Green(4/9))*	
	4 - Sufficient nutritious **water, forage, and food stores** available (inside and/or outside the hive), and young brood being fed?	
	5 - No (apparent) stressors present that would lead to reduced colony survival and/or growth potential?***	
	6 - Suitable space (not too much or too little) for current & near-term expected colony size that is sanitary, defendable, and room for egg laying?	

* √ = Yes; X = No; na = Not Assessed; Use "?" only if unsure and follow up as needed.

** Including: feeding brood, caring for queen, thermoregulation, foraging, house cleaning, undertaking, guarding.

*** If unsure, follow up with more detailed assessment as soon as possible.

General Notes and Observations

Current hive weight (lb / kg): _____; Change from last measure (lb / kg): _____

v2018Feb26

Form comments and questions to dick.rogers@bayer.com

ANEXO B – MÉTRICAS DE AVALIAÇÃO DO DESEMPENHO DOS ALGORITMOS

B.1 Métricas de avaliação da classificação

Para avaliar o desempenho dos algoritmos de classificação, foram utilizadas seis métricas: (i) *Accuracy*, (ii) *Precision*, (iii) *Recall* ou *sensitivity*, (iv) *F1-score*, (v) *Area Under Receiver Operating Characteristic Curve* (AUC ROC) e (vi) *Log loss*, todas definidos abaixo. Nas equações a seguir, tp = verdadeiro positivo, tn = verdadeiro negativo, fp = falso positivo e fn = falso negativo.

(i) *Accuracy* é a acurácia ou exatidão geral do modelo de classificação, que pode ser calculado através da expressão (B.1).

$$CA = \frac{vp + vn}{tp + tn + fp + fn} \tag{B.1}$$

(ii) *Precision* expressa a proporção de amostras classificadas corretamente, considerando o conjunto de todas as amostras classificadas (correta e incorretamente). Valores próximos a 1 significam que o algoritmo retorna mais resultados relevantes que irrelevantes. Pode ser calculado pela expressão (B.2).

$$precision = \frac{tp}{tp + fp} \tag{B.2}$$

(iii) *Recall* ou *sensitivity* ou taxa de verdadeiro positivo explica a eficácia com que o classificador identifica previsões positivas, ou seja, a capacidade do modelo de identificar corretamente quais amostras pertencem a uma determinada classe; calculado por (B.3).

$$recall = \frac{tp}{tp + fn} \tag{B.3}$$

(iv) *F1-score* é uma maneira de equilibrar a *precision* e a *sensitivity*, mas sem sofrer com o problema que a precisão sofre quando há um enorme desequilíbrio de classe. É a média harmônica entre *precision* e a *sensitivity* (B.4).

$$F1\text{-}score = 2 * \frac{precision * recall}{precision * recall} \tag{B.4}$$

(v) *Area Under Receiver Operating Characteristic Curve* (ROC AUC) refere-se à probabilidade de um classificador classificar positivamente uma instância escolhida aleatoriamente. Um valor igual a 1 significa um classificador perfeito, ou seja, que prediz todas as amostras de teste, um valor abaixo de 0,5 significa um modelo ineficaz (B.5) onde $x = 1 - recall$, chamado também de *specificity* ou taxa de verdadeiro negativo.

$$AUC = \int_0^1 recall(x)\, dx \tag{B.5}$$

(vi) *Log loss* ou *binary cross-entropy* é outra métrica baseada em probabilidades. Ela mede a incerteza das probabilidades do modelo comparando-as com os rótulos verdadeiros. Um valor baixo de *Log loss* significa melhores previsões. A função de custo para calcular a *Log loss* é a equação B.6.

$$logLoss = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p(y_i)) + (1 - y_i) * log(1 - p(y_i)) \tag{B.6}$$

Em B.6, y é uma dada classe e $p(y)$ é a probabilidade de uma amostra ser da classe especificada para todos os N pontos.

B.2 Métricas de avaliação da regressão

Para avaliar a qualidade do modelo proposto baseado em regressão, foram utilizadas duas métricas: (i) Erro Quadrático Médio (MSE) e a (ii) Raiz do Erro Quadrático Médio (RMSE), definido como:

(i) Erro Quadrático Médio (MSE) mede a diferença quadrática média entre os erros entre os valores estimados e os valores reais, calculada por B.7.

$$MSE = \frac{1}{N}\sum_{t=1}^{N} e_t^2 \tag{B.7}$$

(ii) Raiz do Erro Quadrático Médio (RMSE) mede a diferença quadrática média entre os valores reais e estimados, pode ser calculada por B.8.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (e_i - \hat{e}_i)^2} \tag{B.8}$$

Em B.7 e B.8, e representa o erro entre um valor estimado e um valor real. Valores baixos de MSE e RMSE indicam um bom modelo de predição.