



UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA
MESTRADO ACADÊMICO EM ENGENHARIA DE TELEINFORMÁTICA

ÍCARO DE LIMA RODRIGUES

**PADRÕES BIOACÚSTICOS COMO IDENTIFICADORES PRECISOS DA PRESENÇA
DE ABELHA RAINHA EM COLMEIAS DE ABELHAS COM E SEM FERRÃO**

FORTALEZA

2023

ÍCARO DE LIMA RODRIGUES

PADRÕES BIOACÚSTICOS COMO IDENTIFICADORES PRECISOS DA PRESENÇA DE
ABELHA RAINHA EM COLMEIAS DE ABELHAS COM E SEM FERRÃO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do 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 mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Danielo Gonçalves Gomes

FORTALEZA

2023

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

R617p Rodrigues, Ícaro de Lima.

Padrões Bioacústicos como identificadores precisos da presença de abelha rainha em colmeias de abelhas com e sem ferrão / Ícaro de Lima Rodrigues. – 2023.
75 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2023.

Orientação: Prof. Dr. Danielo Gonçalves Gomes.

1. Processamento de áudio. 2. Aprendizado de máquina. 3. Abelhas com e sem ferrão. 4. Ciência de dados. 5. Bioacústica. I. Título.

CDD 621.38

ÍCARO DE LIMA RODRIGUES

PADRÕES BIOACÚSTICOS COMO IDENTIFICADORES PRECISOS DA PRESENÇA DE
ABELHA RAINHA EM COLMEIAS DE ABELHAS COM E SEM FERRÃO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do 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 mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em:

BANCA EXAMINADORA

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

Dr. Charles Fernando dos Santos
+abelhas

Dr. Guilherme Gomes
Predikta Soluções em Pesquisa LTDA

Prof. Dr. Luis Ernesto Arruda Bezerra
Instituto de Ciências do Mar (LABOMAR)

À minha família, por servirem como a base mais firme de sustentação, motivação e perseverança na minha vida. Por me ajudarem a trilhar todos os caminhos. Por estarem sempre comigo para dividir cada conquista nossa. Vocês são e sempre serão o meu bem mais importante e precioso.

AGRADECIMENTOS

Aos meus pais, Cícero e Maêda, que se fizeram sempre presentes e sempre ao meu lado servindo de base para cada passo desta caminhada. É uma honra poder dividir cada conquista com vocês. Vocês são meu bem mais precioso e o motivo de sempre ir em frente.

Aos meus irmãos, Iarli e Victor, que sempre foram meus parceiros e dividiram cada momento nessa jornada comigo. Obrigado por todas as palavras de apoio, momentos de ajuda, risadas, companheirismo, etc. Vocês são partes fundamentais da minha vida.

Aos meus avós, José e Marieta, Francisco Izidorio e Maria, que me ensinaram sobre a importância dos estudos e sobre o valor da honestidade. Por me ensinar valores que eu nunca esqueci e que sempre me guiaram por todos os momentos. Também a todos os membros da minha família Lima e minha família Rodrigues. Me sinto honrado de fazer parte dessas famílias.

À minha namorada Camila que durante toda nossa trajetória, desde o ensino médio, se dedicou e acreditou em mim em momentos que nem mesmo eu acreditava. Obrigado por estar comigo e por ser tão importante na minha vida.

A todos meus primos e amigos, em especial Felipe, Lucas, Laenne, Caio, Juliana, L. Teixeira, Paulo Ricardo, Samuel, Keila, Alex, Israel, Larissa, Moita e Bia. Obrigado por todo o suporte, momentos e alegrias compartilhados. Vocês foram fundamentais em toda esta trilha.

Aos meus amigos da UFC, IFCE e do *Felaship* que levo para toda vida, em especial Heitor, Leonel, André, Jefferson, Stefan, Tavares, Paulo, Simão, Amanda, Silvia, Davi, Diego, David, Rodrigo, entre outros. Vocês são incríveis. Muito obrigado!

Agradeço aos amigos e professores do projeto *Sm@rtbee*. Por todo o conhecimento compartilhado, por todo apoio e paciência com cada integrante. Sem o projeto, este trabalho jamais seria possível. Nominalmente agradeço aos professores Danielo, Breno e Rafael e também ao Davyd, Isac, Daniel, Davi, Marcos Augusto, Carolina, entre outros amigos do projeto.

Agradeço a todos os professores, coordenação e secretaria do PPGETI. Por toda atenção, disponibilidade e dedicação para ensinar e ajudar os alunos. Em especial ao meu orientador Prof. Dr. Danielo Gonçalves Gomes por toda a amizade, empenho e paciência na orientação deste e de tantos outros trabalhos e projetos. Por toda a confiança depositada em mim desde a graduação.

Agradeço em especial ao aluno da UFRS Alex Otesbelgue, pela parceria, colaboração, amizade e também pelo coleta dos dados e experimento de abelhas sem ferrão que foi cedido por ele para esta dissertação.

Agradeço aos membros da banca Dr. Charles Fernando dos Santos, Dr. Guilherme Gomes e Prof. Dr. Luis Ernesto Arruda Bezerra por aceitarem o convite para participação nesta banca além de suas valiosas contribuições.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento da pesquisa de mestrado via bolsa de estudos.

“E no final, assim calado, eu sei que vou ser coroadado rei de mim”

(De onde vem a calma - Los Hermanos)

RESUMO

Estima-se que 35% das culturas agrícolas e quase 90% das plantas silvestres com flores dependem de agentes polinizadores. Dentre os insetos polinizadores, as abelhas são os principais e mais importantes. Ainda assim, cerca de 40% das espécies de abelhas no mundo todo estão em declínio. Para monitorar suas colônias e evitar riscos ao bem-estar delas, os apicultores fazem checagens periódicas, tradicionalmente feitas de forma manual e invasiva através da abertura da colmeia. Um dos riscos que pode levar ao colapso e até mesmo a morte da colônia é a perda da abelha rainha. Uma forma não-invasiva de monitoramento remoto e em tempo real é através de microfones e sensores de áudio que detectam os padrões bioacústicos da colônia (e.g. bater de asas, vibrações). Pensando nisso, neste trabalho realizamos experimentos com o áudio produzido por duas espécies diferentes de abelhas: abelhas melíferas africanizadas (com ferrão) e abelhas jataí (sem ferrão) para identificar a presença/ausência de rainha. Ambas as espécies são muito presentes no Brasil. O primeiro experimento possui abordagem de monitoramento em tempo real e fluxo de dados e por isso foram aplicados classificadores incrementais para avaliação de desempenho e validação desta abordagem. Os classificadores incrementais utilizados foram o *Naive Bayes*, o *Hoeffding Tree* e o *Adaptive Random Forest*. No segundo experimento, foram utilizados classificadores *in batch* para classificação entre múltiplas colmeias para determinar quais possuem presença de rainha e em quais a rainha está ausente. Os classificadores *in batch* utilizados foram o *Multilayer Perceptron*, a *Extreme Learning Machine* e o *AdaBoost*. Em ambos os experimentos, os classificadores passaram por análise de desempenho comparando as métricas de avaliação tempo de execução e acurácia. Como resultado, obtivemos que para o caso de classificadores incrementais e fluxo de dados, o classificador *Naive Bayes* obteve o melhor desempenho. Além disso, concluímos que utilizar 10 janelas de 1 segundo para amostragem diária do áudio é o suficiente para detectar a presença de rainha, sendo bem menos custoso computacionalmente. Já para o caso dos classificadores *in batch*, o MLP obteve um melhor desempenho. Finalmente, o monitoramento em tempo real de uma colônia se mostrou a técnica mais acurada e precisa do que uma abordagem entre colônias usando lotes de dados.

Palavras-chave: Abelhas com e sem ferrão. Bioacústica. Aprendizado de máquina. Ciência de dados.

ABSTRACT

It is estimated that 35% of agriculture crops and almost 90% of wild flower plants depend on pollinating agents. Among the pollinating insects bees are the main and most important. However about 40% of the world's bee species are dying. To monitor their colonies and avoid risks to their well-being, beekeepers do periodic checks traditionally done manually and invasively by opening the hive. A risk that can lead to the collapse and even the death of the colony is the loss of the queen bee. A non-invasive way of remote, real-time monitoring is through microphones and audio sensors that detect the colony's bioacoustic patterns (e.g. wing flapping, vibrations). With that in mind, in this work we performed experiments with the audio produced by two different bee species: africanized honey bees (sting) and jataí bees (stingless) to identify the presence/absence of a queen. Both species with widely presence in Brazil. The first experiment has a real-time monitoring approach and datastreams and therefore incremental classifiers were applied for performance evaluation and validation of this approach. The incremental classifiers used were Naive Bayes, Hoeffding Tree and Adaptive Random Forest. In the second experiment, in batch classifiers were used to classify among multiple hives to determine which ones have queen presence and which ones have queen absence. The classifiers in batch used were Multilayer Perceptron, Extreme Learning Machine and AdaBoost. In both experiments, the classifiers underwent performance evaluation comparing the metrics of execution time and accuracy. As result, the Naive Bayes classifier obtained the best performance for the case of incremental classifiers and datastreams. Furthermore we conclude that using 10 windows of 1 second for daily audio sampling is enough to detect the queen presence and is much less expensive computationally. For the case of in batch classifiers, MLP obtained a better performance. Finally, the colony real-time monitoring was also more accurate and precise than a multiple cross-colony approach using data batches.

Keywords: Honey and stingless bees. Bioacoustics. Machine Learning. Data science.

LISTA DE FIGURAS

Figura 1 – Objetivos de Desenvolvimento Sustentável.	19
Figura 2 – Diagrama com uma visão geral da metodologia deste trabalho.	22
Figura 3 – Ilustração do funcionamento de uma árvore de decisão simples.	30
Figura 4 – Ilustração do funcionamento de uma floresta aleatória.	32
Figura 5 – Ilustração da arquitetura básica de um MLP.	34
Figura 6 – Colmeia de observação para captura dos áudios de abelhas melíferas.	42
Figura 7 – Esquema para captura dos áudios de abelhas sem ferrão.	42
Figura 8 – Espectrograma de um intervalo de alguns segundos do último dia com rainha (esquerda) e o dia sem rainha (direita).	44
Figura 9 – Espectrograma de um intervalo de alguns segundos de uma das colônias com rainha (esquerda) e uma sem rainha (direita).	46
Figura 10 – Comparativo do tempo de execução do algoritmo <i>Hoeffding Tree</i> (HT).	51
Figura 11 – Comparativo do tempo de execução do algoritmo <i>Naive Bayes</i> (NB).	53
Figura 12 – Comparativo do tempo de execução do algoritmo <i>Random Forest</i> (RF).	55
Figura 13 – Tempos de execução entre os classificadores.	56
Figura 14 – Acurácia média entre os classificadores.	57
Figura 15 – Matriz de confusão média para 10 janelas de 1 segundo com NB e HT.	58
Figura 16 – Tempo de execução entre os classificadores.	62
Figura 17 – Acurácia entre os classificadores.	63
Figura 18 – Matriz de confusão média para 200 amostras com MLP e ELM.	64

LISTA DE TABELAS

Tabela 1 – Resultado das métricas (média \pm desvio padrão) de HT.	52
Tabela 2 – Resultado das métricas (média \pm desvio-padrão) de NB.	54
Tabela 3 – Resultado das métricas (média \pm desvio-padrão) de RF.	55
Tabela 4 – Resultado da análise com dois ou seis agrupamentos.	59
Tabela 5 – Resultado da análise com o k-médias com $2 \leq k \leq 6$	59
Tabela 6 – Resultado das métricas do MLP.	60
Tabela 7 – Resultado das métricas do ELM.	61
Tabela 8 – Resultado das métricas do ADB.	61
Tabela 9 – Objetivos (geral e específicos) e evidências dos seus alcances.	69

LISTA DE ABREVIATURAS E SIGLAS

ADB	<i>AdaBoost</i>
AP	Apicultura de Precisão
CH	Calinski-Harabasz
CNN	<i>Convolutional Neural Network</i>
DB	Davies-Bouldin
ELM	<i>Extreme Learning Machine</i>
FFT	<i>Fast Fourier Transform</i>
FN	Falsos Negativos
FP	Falsos Positivos
HT	<i>Hoeffding Tree</i>
IoT	<i>Internet of Things</i> ou Internet das coisas
MFCCs	<i>Mel-frequency cepstral coefficients</i>
MLP	<i>Multilayer Perceptron</i>
NB	<i>Naive Bayes</i>
QP	Questões de Pesquisa
RF	<i>Adaptive Random Forest</i>
S-T	Transformada-S
SLFN	<i>Single-hidden Layer Feedforward Neural Network</i>
STFT	<i>Short-term Fourier Transform</i>
SVD	<i>Single Value Decomposition</i>
SVM	<i>Support Vector Machine</i>
TIC	Tecnologias da Informação e Comunicação
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Contextualização e Justificativa	16
1.2	Caracterização do Problema	18
1.3	Questões de Pesquisa	19
1.3.1	<i>Hipóteses</i>	20
1.4	Objetivos	21
1.5	Visão Geral da Metodologia	21
1.6	Organização da Dissertação	24
2	CONCEITOS E REVISÃO DA LITERATURA	25
2.1	Apicultura de Precisão	25
2.2	Bioacústica	25
2.3	Classificação	27
2.3.1	<i>Classificadores Incrementais</i>	28
2.3.1.1	<i>Naive Bayes (NB)</i>	28
2.3.1.2	<i>Hoeffding Tree (HT)</i>	30
2.3.1.3	<i>Adaptive Random Forest (RF)</i>	31
2.3.2	<i>Classicadores (in batch)</i>	33
2.3.2.1	<i>Multilayer perceptron (MLP)</i>	33
2.3.2.2	<i>Extreme Learning Machine (ELM)</i>	35
2.3.2.3	<i>AdaBoost (ADB)</i>	36
2.4	Outros conceitos	37
2.4.1	<i>k-médias</i>	37
2.4.2	<i>Índices de validação dos agrupamentos</i>	38
2.5	Sumário do capítulo	40
3	MATERIAL E MÉTODOS	41
3.1	Objeto de estudo	41
3.2	Delineamento	43
3.2.1	<i>Abelhas melíferas</i>	44
3.2.2	<i>Abelhas jataí</i>	45
3.3	Avaliação de desempenho dos classificadores	47

3.4	Testes estatísticos	49
3.5	Sumário do capítulo	50
4	RESULTADOS	51
4.1	Resultados do experimento com a colônia de abelhas com ferrão (<i>Apis mellifera</i>)	51
4.1.1	<i>Hoeffding Tree</i>	51
4.1.2	<i>Naive Bayes</i>	53
4.1.3	<i>Adaptive Random Forest</i>	54
4.1.4	<i>Comparando os 3 classificadores</i>	56
4.2	Resultados do experimento com colônias de abelhas sem ferrão (<i>Tetragonisca fiebrigi</i>)	58
4.2.1	<i>Multilayer Perceptron</i>	60
4.2.2	<i>Extreme Learning Machine</i>	60
4.2.3	<i>AdaBoost</i>	61
4.2.4	<i>Comparando os 3 classificadores</i>	62
4.3	Discussão dos resultados	64
4.4	Sumário do Capítulo	66
5	CONCLUSÃO	68
5.1	Contribuições e Objetivos Alcançados	68
5.2	Publicações	69
5.3	Algumas Perspectivas	70
	REFERÊNCIAS	71

1 INTRODUÇÃO

Este capítulo contextualiza esta dissertação. Aqui apresentamos o problema abordado, justificativa e motivação para mitigá-lo bem como o objetivo central desta pesquisa. Primeiramente, na Seção 1.1, apresentamos o contexto e justificativa desta pesquisa. Depois, na Seção 1.2 abordamos a problemática deste estudo. A seguir, levantamos quatro questões de pesquisa (Seção 1.3) e suas respectivas hipóteses de resposta (Seção 1.3.1). Na sequência, a Seção 1.4 traz os objetivos (gerais e específicos) e uma síntese da metodologia geral é discorrida na Seção 1.5. Concluímos este capítulo descrevendo a estrutura da dissertação na Seção 1.6.

1.1 Contextualização e Justificativa

Estima-se que cerca de 35% das culturas agrícolas e quase 90% das plantas silvestres com flores dependam de agentes polinizadores (KLEIN *et al.*, 2007; OLLERTON *et al.*, 2011). A polinização para produção comercial de alimentos em escala mundial é de centenas de bilhões de dólares (LAUTENBACH *et al.*, 2012). Além disso, os polinizadores são diretamente ligados ao bem-estar humano através da manutenção do ecossistema, reprodução de plantas, produção de safras, etc. (POTTS *et al.*, 2016). Dentre os insetos polinizadores, as abelhas são os principais e mais importantes (BROWN *et al.*, 2016). Entretanto, estima-se que cerca de 40% das espécies de abelhas em todo o mundo estão em declínio (SÁNCHEZ-BAYO; WYCKHUYS, 2019).

A fim de monitorar suas colônias e evitar anomalias (e.g. pragas, doenças, enxameamento, perda da rainha) os apicultores fazem uma checagem periódica, atestando o bem-estar de cada colônia. Contudo, este tipo de inspeção é tradicionalmente realizada *in loco* de forma manual e invasiva através do deslocamento do apicultor e abertura das colmeias. Diante disso ressalta-se a importância da Apicultura de Precisão (AP). A AP é um ramo da Agricultura de Precisão que define estratégias de gerenciamento dos apiários baseado no monitoramento das colônias de forma a reduzir o consumo de recursos e maximizar a produtividade (ZACEPINS *et al.*, 2015).

Uma das condições para uma colônia de abelhas em pleno bem-estar e forte é ter uma rainha jovem e saudável. Se esta condição não for satisfeita, pode representar sérios riscos ao bem-estar da colônia, podendo até mesmo levar ao colapso e à morte. A ausência da rainha quebra a coesão da colônia devido à falta dos feromônios, leva à descontinuidade da produção de crias e desestrutura o trabalho normal das operárias. As abelhas têm a capacidade de reconhecer

rapidamente a ausência de sua rainha e é recomendável que o apicultor disponha de meios para também fazê-lo, desta forma, o apicultor pode tomar decisões que evitem ou que contornem a situação da colônia (WINSTON, 1991).

Nos últimos anos, pesquisadores têm utilizado de métodos não-invasivos de monitoramento remoto e em tempo real, através da AP, para gerenciamento de apiários e colmeias a fim de ajudar apicultores a entender o que está acontecendo dentro das colmeias sem precisar abri-las (DINEVA *et al.*, 2018; MEIKLE *et al.*, 2017). Atenção especial tem sido dada ao monitoramento de variáveis físicas da colmeia, e.g. temperatura e umidade (BRAGA *et al.*, 2021), peso (SILVA *et al.*, 2020; BRAGA *et al.*, 2020), gases (WILK *et al.*, 2021), padrões de áudio (RODRIGUES *et al.*, 2022).

Particularmente com relação ao áudio (zumbido) das abelhas, um algoritmo de reconhecimento de padrões bioacústicos pode ser treinado para identificar adversidades em uma colônia, tais como enxameamento (ŽGANK, 2018), infestação de parasitas (HADJUR *et al.*, 2022) e ausência de rainha (HOWARD *et al.*, 2013). A título ilustrativo e prático, o zumbido das abelhas difere quando a rainha está presente ou ausente¹ até mesmo para ouvido humano. O som produzido pelas abelhas é caracterizado pelo conjunto de diversas atividades produzidas por cada uma dentro da colmeia tais como vibrações, bater de asas, movimentações dentro da colmeia.

Algumas soluções tem buscado desenvolver sistemas em tempo real (*online*), com arquiteturas de *Internet of Things* ou Internet das coisas (IoT) (ZGANK, 2020) para um melhor suporte e eficiência para o apicultor. Entretanto, o monitoramento em tempo real de uma colônia demanda classificadores incrementais ou do tipo *online*. Esses tipos de classificadores recebem, avaliam e se adaptam aos dados em tempo de execução sem a necessidade de *datasets* prévios para treinamento, como no caso dos classificadores *in batch* (*offline*). Além disso, grandes massas de dados, como áudio, são computacionalmente custosas tanto em processamento quanto em tempo de transmissão pela rede. Este custo computacional pode atrasar uma tomada de decisão do apicultor trazendo riscos ao bem-estar das colônias. Portanto, há necessidade de encontrar uma forma eficaz de identificar a presença de rainha na colônia utilizando uma taxa de gravação diária aliado a classificadores *online*.

Enquanto isso, o monitoramento entre múltiplas colônias sugere a utilização de modelos convencionais (*in batch*) de aprendizado de máquina. Geralmente os dados provenientes deste tipo de análise são balanceados e se busca modelos que generalizem bem as diferenças entre

¹ <https://smartbee-ufc.github.io/bioacoustics/>

as classes mesmo lidando com múltiplas colmeias, cada qual com suas próprias características individuais (e.g. atividade na colmeia, número de forrageiras, quantidade de operárias, cria, alimento).

1.2 Caracterização do Problema

Assim, definimos a problemática desta dissertação da seguinte forma:

Como identificar a presença de abelha rainha de forma rápida, acurada e precisa através dos padrões bioacústicos de abelhas, em especial, de abelhas comuns no ecossistema brasileiro?

Infelizmente, pouco se tem utilizado de técnicas de monitoramento e aprendizado de máquina usando padrões bioacústicos de espécies de abelhas comuns no Brasil. A grande maioria dos estudos recentes abordam as abelhas melíferas (*Apis mellifera*) europeias em clima temperado (ABDOLLAHI *et al.*, 2022). Uma pequena parte até estuda abelhas sem ferrão, mas em nenhum estudo se abordou padrões bioacústicos de abelhas jataí, estudadas neste trabalho, por exemplo, comuns no Sul do Brasil. Bem como nenhum estudo abordou os padrões acústicos de abelhas melíferas do tipo africanizadas, que são comuns no Nordeste. Logo, poucos trabalhos recentes tem tratado destas questões voltadas ao bem-estar de abelhas comuns no ecossistema brasileiro. Sendo o Brasil um país continental e com uma grande biodiversidade, urge a necessidade de mais dados provenientes de espécies adaptadas aos climas e regiões diferentes do território nacional. Estas técnicas e dados podem auxiliar apicultores e meliponicultores do país a monitorar de forma não-invasiva, rápida e eficiente suas colônias de abelhas, garantindo o bem-estar delas.

Além disso, a utilização de técnicas diferentes de classificação podem também ajudar no manejo de colmeias sensoriadas. A aplicação de novas técnicas como a classificação incremental, que é voltado para lidar com fluxos de dados, para definir, comparar e determinar se esta é, por exemplo, uma melhor forma de utilizar os padrões bioacústicos para identificar de maneira eficiente a presença de rainha na colônia. Atualmente, no estado da arte para classificação e monitoramento de colmeias não se tem utilizado classificadores incrementais e fluxos de dados. Portanto, além de utilizar duas espécies de abelhas com e sem ferrão comuns em regiões diferentes do Brasil, buscando abranger a realidade de diversos tipos de apicultores/meliponicultores do país, também serão comparados os classificadores convencionais com classificadores incrementais, respeitando o tipo de delineamento experimental voltado para cada um destes tipos de classificadores.

Ademais, considerando os 17 Objetivos de Desenvolvimento Sustentável (ODS) estabelecidos pelas Nações Unidas (Figura 1), esta dissertação elucida alguns potenciais impactos econômicos e ambientais em alguns dos ODS. Mais especificamente nos ODS número 2 e número 11. Primeiramente, no ODS número 2 - Fome zero e agricultura sustentável: existe o forte impacto econômico e ambiental dos serviços de polinização das abelhas na agricultura. A garantia do bem-estar das abelhas, evitando quaisquer tipos de riscos que possam levar as colônias ao colapso (e.g. ausência de rainha) de forma não-invasiva, são formas de promover e garantir uma agricultura sustentável. Com respeito a ODS número 11 - Cidades e comunidades sustentáveis: destacamos o impacto positivo que as abelhas possuem no ecossistema, afetando as plantas, outros animais e nós seres humanos. Além disso, a utilização de um sistema de monitoramento remoto em tempo real através de fluxo de dados de redes de sensores significa avanços tecnológicos para a comunidade enquanto busca um desenvolvimento sustentável, evitando prejuízos às abelhas ou ao meio ambiente. Permitindo assim o crescimento econômico e ambiental sustentável das comunidades que são afetadas direta ou indiretamente pelos serviços de polinização das abelhas.

Figura 1 – Objetivos de Desenvolvimento Sustentável.



Fonte: <https://brasil.un.org/pt-br/sdgs>

1.3 Questões de Pesquisa

Considerando a problemática deste estudo, levantamos quatro Questões de Pesquisa (QP) na qualidade de pontos de partida desta dissertação:

QP #1: *É possível identificar a presença de rainha com dados acústicos monitorados*

em tempo real?

QP #2: *Qual a melhor taxa de amostragem bioacústica diária para um sistema de monitoramento de uma colônia?*

QP #3: *É possível distinguir colônias com rainha de colônias sem rainha em cenários com múltiplas colônias?*

QP #4: *Qual a melhor estratégia de monitoramento: em tempo real de cada colônia individualmente ou em lotes de dados de múltiplas colônias?*

Para responder cada uma das quatro perguntas, foram utilizados classificadores incrementais e *in batch*, para os casos de monitoramento em tempo real e monitoramento entre colônias, respectivamente, de espécies de abelhas comuns no ecossistema brasileiro. Além disso foram realizadas análises de desempenho para avaliar a influência da taxa de amostragem diária do áudio produzido pelas colônias no tempo de execução e na acurácia dos classificadores. Além da comparação entre os classificadores para avaliar os que obtiveram melhores desempenhos.

1.3.1 Hipóteses

A **Hipótese principal** desta dissertação é que *através dos padrões bioacústicos de abelhas comuns no ecossistema brasileiro é possível identificar a presença de abelha rainha de forma rápida, acurada e precisa*. Acredita-se que análises de padrões bioacústicos da literatura para abelhas de clima temperado são passíveis de aplicação para detectar a presença de rainha em abelhas comuns no ecossistema brasileiro de forma rápida, acurada e precisa. Além desta, podemos elencar 4 hipóteses específicas as quais respondem provisoriamente cada QP:

Hipótese #1 relacionada à **QP #1**: a primeira hipótese específica é que seja possível utilizar classificadores incrementais que lidam com fluxo de dados para assim identificar a presença de rainha através do monitoramento de áudio em tempo real;

Hipótese #2 relacionada à **QP #2**: a segunda hipótese específica é que seja possível determinar uma taxa de amostragem diária para monitoramentos em tempo real de forma que o custo de tempo seja reduzido o máximo possível sem prejudicar a eficiência dos classificadores;

Hipótese #3 relacionada à **QP #3**: a terceira hipótese específica é que o padrão bioacústico relativo à ausência de rainha é superior e mais relevante mesmo em cenários de múltiplas colônias onde cada qual tem suas individualidades (e.g. atividade forrageira, quantidade de abelhas adultas, cria, alimento);

Hipótese #4 relacionada à **QP #4**: a quarta hipótese específica é que seja possível

recomendar a melhor estratégia para monitoramento de colmeias de abelhas entre o monitoramento em tempo real individualmente ou o monitoramento através de lotes de dados de múltiplas colmeias.

1.4 Objetivos

O **Objetivo geral** deste trabalho é *detectar a presença da rainha em colmeias de abelhas de espécies com e sem ferrão, comuns no ecossistema brasileiro, através de seus padrões bioacústicos de forma rápida, acurada e precisa*. Este objetivo está diretamente associado a hipótese principal deste estudo.

Além disso, os objetivos específicos desta dissertação que estão ligados com as hipóteses específicas e conseqüentemente às perguntas de pesquisa são relacionados a seguir.

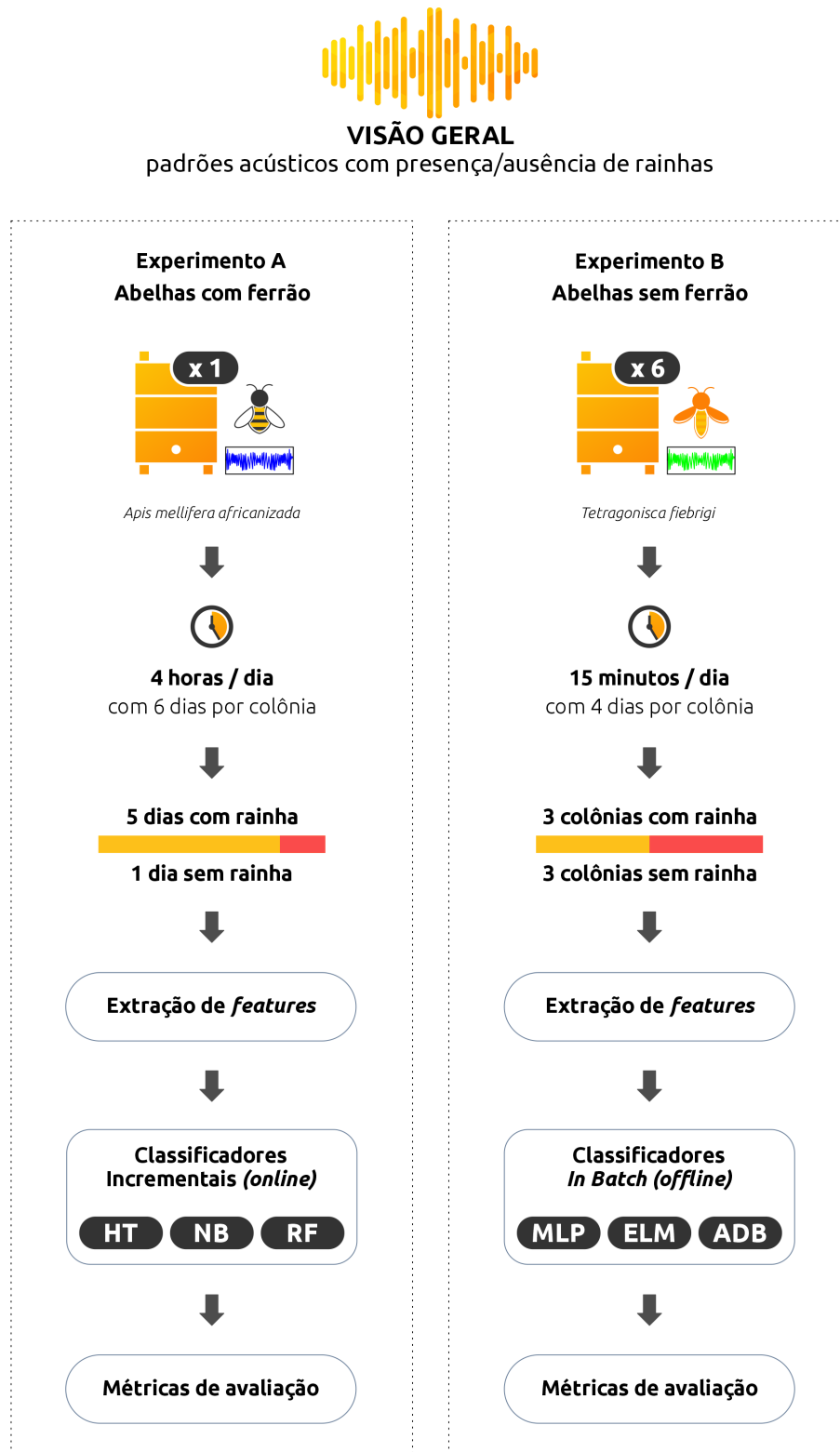
- **Objetivo específico #1:** Validar a aplicação de classificadores incrementais para fluxo de dados de áudio seguindo um monitoramento em tempo real.
- **Objetivo específico #2:** Determinar uma taxa de amostragem diária ideal para monitoramento periódico e individual de colônias de abelhas.
- **Objetivo específico #3:** Identificar através de análise de agrupamentos que os padrões acústicos das abelhas são mais relevantes no que diz respeito a presença de rainha, do que as características de "assinatura" individuais de cada uma das colônias.
- **Objetivo específico #4:** Recomendar uma melhor estratégia de monitoramento para dados bioacústicos de colônias de abelhas.

1.5 Visão Geral da Metodologia

Na Figura 2 temos uma síntese visual do procedimento adotado como metodologia para esta dissertação. De maneira geral temos um mesmo processo para ambos os experimentos. Como passo inicial, temos a gravação do áudio dentro das colmeias, cada experimento com sua própria taxa de amostragem e balanceamento das classes. O áudio passa por um processamento para extração de *features* e em seguida essas características extraídas são passadas por três classificadores para cada caso (*online* e *offline* ou incremental e *in batch*) e em seguida obtem-se na saída de cada classificador as suas métricas de avaliação (e.g. acurácia, precisão, *recall* e *f-score*).

Aprofundando um pouco mais em cada experimento, o primeiro experimento foi

Figura 2 – Diagrama com uma visão geral da metodologia deste trabalho.



Fonte: elaborado pelo autor (2021).

realizados com abelhas melíferas (*Apis mellifera*) do tipo africanizadas, localizadas em um apiário em Fortaleza-CE. Este experimento foi feito pensando em um monitoramento em tempo real de uma colônia. Todo dia foram gravados aproximadamente 4 horas de gravação durante 6 dias. Essa configuração do experimento aproxima-se de um fluxo de dados ou *datastream*. Por consequência, o tipo de classificação preparada para lidar com fluxo de dados é a incremental ou *online*. Dos 6 dias de gravação, em 5 o funcionamento da colônia estava normal e no último dia a rainha foi removida da colônia para simular uma perda súbita da rainha (e.g. doença, morte por predadores). Os dados não são balanceados mas é algo que é mais próximo da realidade e do fluxo de dados. Se espera que durante um monitoramento de um sistema na maior parte do tempo ele apresente um funcionamento normal. Logo, o classificador incremental tem que poder lidar com este tipo de desbalanceamento. Em seguida, temos a extração de característica ou *features*, utilizando as *Mel-frequency cepstral coefficients* (MFCCs) e métricas e estatísticas do sinal (e.g. energia, centróide espectral, fluxo espectral). Depois de extraídas as características, o conjunto de dados é levado aos três classificadores incrementais utilizados neste trabalho: *Naive Bayes* (NB), *Hoeffding Tree* (HT) e *Adaptive Random Forest* (RF). A saída produzida por cada classificador geram as métricas de avaliação de cada um. Essas métricas são utilizadas para avaliar e comparar o desempenho de cada classificador incremental.

No segundo experimento, temos um processo muito parecido com o primeiro. O segundo experimento foi realizados com abelhas jataí (*Tetragonisca fiebrigi*) localizadas em um meliponário em Viamão-RS. Este experimento foi feito com intenção de monitoramento de múltiplas colônias e na comparação entre elas já que metade tinham rainhas presentes e a outra metade não. Foram utilizadas 6 colônias, e as gravações foram de 15 minutos por dia para cada colônia durante um total de 4 dias. Dado este delineamento, as classes são bem balanceadas e portanto é um caso mais adequado para os classificadores *in batch* ou em lote ou ainda *offline*. Após esta etapa, os dados de áudio puros passam por uma etapa de extração de *features* onde novamente se utilizam as MFCCs além de métricas e estatísticas do sinal. Depois de extraídas as características, o conjunto de dados gerado é levado para os três classificadores *in batch* utilizados neste trabalho: *Multilayer Perceptron* (MLP), *Extreme Learning Machine* (ELM) e *AdaBoost* (ADB). Da mesma forma que no primeiro experimento, a saída são as métricas de avaliação que serão usadas para avaliar e comparar o desempenho dos classificadores *in batch* utilizados.

1.6 Organização da Dissertação

Esta dissertação está dividida em cinco capítulos principais. O primeiro capítulo é este apresentado até aqui como a introdução deste trabalho. Os próximos capítulos são apresentados a seguir.

- **Capítulo 2:** aborda os principais conceitos teóricos que estão envolvidos na pesquisa e desenvolvimento desta dissertação. São fundamentados temas como apicultura de precisão e as técnicas de aprendizado de máquina e análise de dados utilizadas. Além de uma breve revisão da literatura.
- **Capítulo 3:** apresenta a descrição e detalhes dos procedimentos que foram realizados com o propósito de atingir os resultados que sustentem os objetivos apresentados na Seção 1.4;
- **Capítulo 4:** descreve os resultados mais importantes para este trabalho que foram obtidos através dos procedimentos descritos no Capítulo 3.
- **Capítulo 5:** apresenta uma síntese final dos resultados obtidos e considerações finais, como as conclusões e contribuições deste trabalho, na perspectiva das QPs (Seção 1.3) e objetivos (Seção 1.4).

2 CONCEITOS E REVISÃO DA LITERATURA

Neste capítulo, apresentamos alguns conceitos úteis para uma boa compreensão da dissertação. Destacamos alguns trabalhos relevantes e recentes sobre Apicultura de Precisão (Seção 2.1), com foco no monitoramento de áudio e padrões bioacústicos das abelhas (Seção 2.2). Em seguida, na Seção 2.3, discutimos brevemente sobre classificação incremental e classificação *in batch* assim como estão brevemente apresentados e discutidos cada um dos classificadores utilizado neste trabalho. Ao final, na Seção 2.4, encontram-se outros métodos de aprendizado de máquina ou métricas relacionadas mas que não fazem parte do escopo de classificação e consequentemente não fazem parte diretamente da análise de desempenho principal desta dissertação.

2.1 Apicultura de Precisão

Apicultura de Precisão (AP) é um ramo da Agricultura de Precisão que trata de estratégias de gerenciamento de apiários com base em monitoramento das colmeias para reduzir o consumo de recursos e maximizar a produtividade das abelhas (ZACEPINS *et al.*, 2015). AP envolve, basicamente, três etapas: coleta de dados, análise de dados e a consequente tomada de decisão no contexto do gerenciamento do apiário (DINEVA *et al.*, 2018). De forma a implementar com sucesso estas 3 etapas, a AP faz uso de Tecnologias da Informação e Comunicação (TIC), tais como sensoriamento embarcado, IoT, Computação em Nuvem, *Big Data* e Aprendizado de Máquina. Através destas tecnologias, é possível coletar uma grande quantidade de dados remotamente e em tempo real para que possam ser analisados em tempo hábil e útil aos manejos dos apicultores no sentido de manter a saúde e bem-estar da colônia. Através da AP, os apiários podem ser monitorados, de maneira não-invasiva e dinâmica para as abelhas, até mesmo durante um inverno severo, quando as inspeções são contraindicadas (MEIKLE *et al.*, 2017).

Dentre os problemas que podem ser tratados pela Apicultura de Precisão, um que é de particular interesse desta dissertação: padrões acústicos que podem indicar riscos ao bem-estar da colmeia, neste caso mais especificamente a ausência de rainha na colônia.

2.2 Bioacústica

A bioacústica diz respeito a uma análise dos padrões acústicos variados produzidos por seres vivos dos mais diversos tipos. Por exemplo, no caso de mamíferos, é possível deter-

minar diferentes espécies de morcegos brasileiros utilizando o padrão acústico do chamado de ecolocalização deles (ARIAS-AGUILAR *et al.*, 2018). No caso de aves, temos por exemplo a utilização de som produzido por pássaros para classificação em larga escala de espécie de pássaros (STOWELL; PLUMBLEY, 2014). Chegando aos insetos, mais especificamente as abelhas, é possível encontrar correlações entre os padrões acústicos emitidos durante a dança entre abelhas forrageiras e a comunicação da localização das reservas de comida (NIEH; ROUBIK, 1998). Não apenas isso, ainda com respeito às abelhas, a intensidade dos sons emitidos em determinadas faixas de frequências aumentam quando uma colônia está a duas ou três semanas de um evento de enxameamento, por exemplo (HUNTER; STEBEL, 2013).

A utilização de padrões acústicos no monitoramento de abelhas tem sido amplamente discutido na literatura por ser uma forma simples (através de uma gama de diversos tipos de microfones) e não-invasiva de monitoramento. De fato, as abelhas podem gerar sinais acústicos específicos através de uma combinação de comportamentos (e.g. movimentação, bater de asas, vibração dos músculos do tórax) (HUNT; RICHARD, 2013; BARTH *et al.*, 2008).

Uma revisão da literatura recente (ABDOLLAHI *et al.*, 2022) utilizou várias bases como *Google Scholar*, *PubMed*, *Scopus*, *IEEE Xplore* e *arXiv*, entre os anos de 2012 e 2021 para encontrar 61 trabalhos que utilizam do monitoramento acústico voltado para abelhas. A maior parte dos trabalhos advém de locais como Estados Unidos e Reino Unido. Dentre os diversos objetivos dos respectivos trabalhos (e.g. detecção de abelhas, força da colônia, exameamento, infestação, ausência de rainha, identificação de espécies), cerca de 15,5% dos trabalhos tratavam da presença/ausência de rainha na colônia. Como é também o objetivo principal deste trabalho.

Além disso, os arquivos de áudio puro podem ter suas características extraídas através de diversas técnicas, como espectrogramas, MFCCs (*Mel-frequency cepstral coefficients*) além de transformadas como a *Fast Fourier Transform* (FFT), *Short-term Fourier Transform* (STFT), Transformada-S (S-T). Dentre todas as técnicas utilizadas, os MFCCs foram de longe o conjunto de características mais populares nos estudos. Por conta disso, neste trabalho foram utilizados também os MFCCs como características principais aliados com algumas características do sinal (e.g. energia, centroide espectral, fluxo espectral) para o treinamento de modelos de aprendizado de máquina para a classificação tradicional (*in batch*) e incremental (*online*).

Focando alguns trabalhos em específico, em (ROBLES-GUERRERO *et al.*, 2017) temos uma análise dos dados acústicos com duas colmeias diferentes, utilizando MFCCs e um modelo de regressão logística para classificação binária utilizando apenas dois coeficientes dos

MFCCs. Já em (NOLASCO *et al.*, 2019) temos a utilização de *Support Vector Machine* (SVM) e *Convolutional Neural Network* (CNN) como modelos de aprendizado de máquina e aprendizado profundo utilizando duas colmeias diferentes. Ao todo, foram 96hs de gravação onde um dia as colmeias passaram com rainha e o outro sem. Além disso, usaram MFCCs entre os métodos para extração de *features* que inclusive alcançou os melhores resultados. Em (ROBLES-GUERRERO *et al.*, 2019) temos novamente os MFCCs e também a utilização de regressão logística para classificação, além de *Single Value Decomposition* (SVD) para auxiliar na visualização. Neste trabalho, os autores lidam com 5 colmeias em dois experimentos espaçados no tempo. Durante os experimentos duas colmeias permaneceram com rainha, uma esteve sempre sem rainha e duas tiveram a rainha removida entre um experimento e outro. O estudo alcança também excelentes resultados e mostra a visualização dos *clusters* de cada colmeia durante o primeiro e segundo experimento. Sendo assim, é possível ver a mudança no padrão dos *clusters* das colmeias que tiveram suas rainhas removidas.

Em nenhum estudo a abordagem de classificação incremental foi utilizado. Os classificadores incrementais são mais adequados para fluxo de dados, ou seja, dados sequenciados e de séries temporais. Desta forma, utilizar classificadores incrementais é uma boa aposta para o caso de monitoramento periódico e em tempo real de colmeias individualmente. Assim como não abordaram abelhas melíferas do tipo africanizadas, e nem mesmo as abelhas-africanas (*Apis mellifera Scutellata*) que tem uma forte proximidade genética com as abelhas do tipo africanizadas estudadas neste trabalho. Além de não mencionarem também as abelhas jataí que são as abelhas sem ferrão estudadas nesta dissertação. Dentre as espécies de abelhas estudadas, a grande maior parte foi de abelhas européias e mamangavas (*bumblebees*). Sendo assim, diferentemente dos trabalhos recente, este estudo se debruça sobre as abelhas comuns no Brasil. Onde a abelha melífera africanizada, que é uma raça híbrida entre a espécie europeia e a africana, é mais comum e mais adaptada ao clima e ambiente do Nordeste brasileiro. Enquanto a abelha jataí é uma espécie sem ferrão mais adaptada e comum no Sul do país.

2.3 Classificação

A classificação é uma das tarefas mais comuns em aprendizado de máquina, também chamado de aprendizado supervisionado. O objetivo da classificação é voltado para determinar ou prever um valor nominal (classe) y , dado um vetor de atributos ou características não rotulados x . Nesta dissertação, abordamos dois tipos de classificadores: (i) incrementais ou

online (Seção 2.3.1) que são voltados para fluxo de dados onde as instâncias dos dados chegam continuamente e periodicamente a medida que são monitorados em tempo real; e os (ii) *in batch* ou *offline* (Seção 2.3.2) voltados para dados em lotes, onde se coleta uma boa quantidade de dados para realizar o treinamento e teste do modelo.

2.3.1 Classificadores Incrementais

Neste trabalho, os classificadores incrementais (*online*) foram utilizados no experimento com abelhas com ferrão, da espécie *Apis mellifera* do tipo africanizada. Como o experimento em questão diz respeito ao monitoramento de uma colônia de abelha por alguns dias corridos em que no último dia ela teve sua rainha removida, a utilização de classificadores incrementais é mais adequada para os dados coletados. As classes são desbalanceadas e os dados são resultados de um monitoramento periódico, ou seja, constituem um fluxo de dados (*data stream*). Os classificadores incrementais são algoritmos de classificação que lidam especificamente com fluxo de dados. Fluxos de dados são sequências de dados contínuos e que evoluem através do tempo. Sendo assim, os classificadores incrementais possuem a tarefa de se atualizar incrementalmente conforme cada amostra do fluxo de dados for chegando (NGUYEN *et al.*, 2015). No caso desta dissertação, os modelos vão classificando e se adaptando a cada amostra de áudio coletada da colônia monitorada, metodologia conhecida como *prequential*. Através da classificação incremental é possível identificar quando a colônia monitorada perde a sua rainha e conseqüentemente é possível gerar alertas para os apicultores responsáveis a fim de auxiliar numa tomada de decisão rápida e eficiente.

2.3.1.1 Naive Bayes (NB)

O classificador *Naive Bayes* (NB) é um classificador probabilístico que é baseado no teorema de Bayes (Equação 2.1). Ele funciona associando a classe mais provável para um dado vetor de atributos de uma amostra. Para isso, ele assume que os atributos das amostras são independentes. Contudo, isto é violado em muitos casos práticos utilizando dados reais, por isso o nome *naive* que quer dizer "ingênuo" em tradução livre. Ainda assim, mesmo sendo uma técnica simples, o *Naive Bayes* é eficiente em diversas aplicações práticas (RISH *et al.*, 2001).

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})} \quad (2.1)$$

Assumindo que um conjunto de dados contém n pontos amostrais \mathbf{x}_i , $i = 1, \dots, n$ com p atributos $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{ip})$. Cada ponto amostral pertence a uma classe $y \in y_1, y_2, \dots, y_m$. O aprendizado acontece com a construção de um modelo probabilístico que associa uma probabilidade posterior de uma classe a um ponto amostral, tal qual: $P(Y = y_j | \mathbf{X} = \mathbf{x}_i)$. Ou em outras palavras, a probabilidade de a classe ser y_j dado uma entrada \mathbf{x}_i . Aplicando no teorema de Bayes nesta perspectiva temos a Equação 2.2.

$$P(y_j | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | y_j)P(y_j)}{P(\mathbf{x}_i)} \quad (2.2)$$

Como já mencionado, assumindo que \mathbf{x}_i são independentes entre si podemos concluir que $P(x_1 | x_2, \dots, x_p, y_j) = P(x_1 | y_j)$ e conseqüentemente chegamos a Equação 2.3.

$$P(\mathbf{x}_i | y_j)P(y_j) = \prod_{k=1}^p P(x_k | y_j)P(y_j) \quad (2.3)$$

E considerando esta igualdade na Equação 2.2 chegamos na Equação 2.4.

$$P(y_j | \mathbf{x}) = \frac{\prod_{k=1}^p P(x_k | y_j)P(y_j)}{P(\mathbf{x})} \quad (2.4)$$

Todavia, é possível notar que o denominador $P(\mathbf{x})$ é fixo para qualquer y , ou seja, não depende da classe. Por esse motivo, esse termo é desconsiderado para classificação, uma vez que só atua como um fator de escala e não influencia na tarefa de classificação. Logo, a classe escolhida é aquela que maximiza o termo numerador $\prod_{k=1}^p P(x_k | y_j)P(y_j)$. O resultado da classe que maximiza o termo numerador é chamada de classe *maximum a posteriori* e é calculada como \hat{y} como mostrado na Equação 2.5.

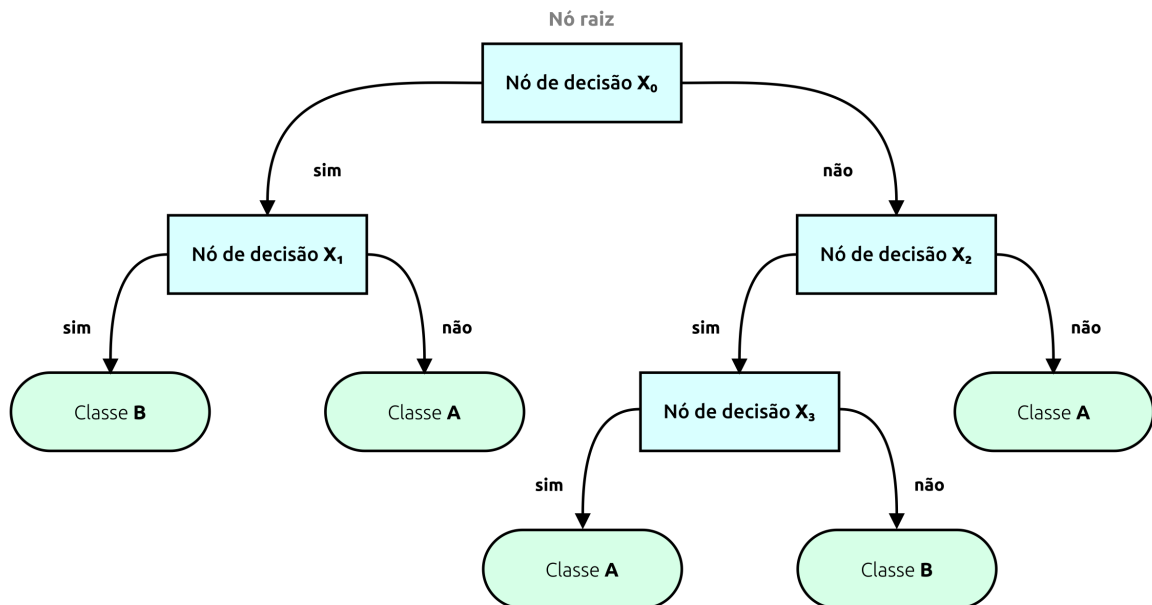
$$\hat{y} = \arg \max_{y_j} \prod_{k=1}^p P(x_k | y_j)P(y_j) \quad (2.5)$$

Sendo assim, é chamado de classificador *Naive Bayes* um modelo que utiliza a Equação 2.5 como critério para classificação (BERRAR, 2018). Para o caso incremental, o algoritmo utilizado aplica o mesmo o classificador NB oriundo do teorema de Bayes para cada entrada \mathbf{x} incidente do fluxo de dados.

2.3.1.2 Hoeffding Tree (HT)

O *Hoeffding Tree* (HT) é um classificador do tipo árvore de decisão adaptado para lidar com fluxo de dados. Árvores de decisão tradicionais definem seus nós de decisão através de múltiplas passagens através dos dados de treino. Entretanto, isso não é possível no caso de um modelo incremental que lida com fluxo de dados. Portanto, o HT é um algoritmo incremental que, dado um novo dado incidente, utiliza os limites de *Hoeffding* para determinar se o melhor nó de decisão é bom suficiente para criar um novo nó no próximo nível da árvore (NGUYEN *et al.*, 2015). Para um melhor entendimento, a Figura 3 ilustra a estrutura geral de uma árvore de decisão. Os nós de decisão verificam as condições de uma nova entrada e direcionam para os nós subsequentes de acordo com o resultado da condição avaliada. As folhas da árvore representam a decisão a se tomar.

Figura 3 – Ilustração do funcionamento de uma árvore de decisão simples.



Fonte: elaborado pelo autor (2023).

Entretanto, conforme mencionado, os nós de decisão são definidos durante a etapa de treinamento, algo que não acontece no caso de classificação incremental. Para tanto é utilizado o limite de *Hoeffding* para resolver este problema. Dadas n observações de uma variável aleatória r com variação R e média observada \bar{r} , o limite de *Hoeffding* garante, com probabilidade $1 - \delta$ que a média real é pelo menos $\bar{r} - \epsilon$, onde ϵ é definido como na equação 2.6.

$$\varepsilon = \sqrt{\frac{R^2 \ln(\frac{1}{\delta})}{2n}} \quad (2.6)$$

Desta forma, sendo $G(X_i)$ a medida heurística para escolher um nó de decisão. Após N observações, numa folha da árvore, seja X_a o nó com a melhor medida heurística e X_b o segundo melhor. Sendo r considerado agora como $r = \Delta G = G(X_a) - G(X_b)$. Aplicando o limite de *Hoeffding*, temos $\bar{r} = \Delta \bar{G} = G(\bar{X}_a) - G(\bar{X}_b)$. E portanto se $\Delta \bar{G} > \varepsilon$ (calculado pela Equação 2.6) é possível afirmar com certo grau de confiança $(1 - \delta)$ que a diferença entre $G(\bar{X}_a)$ e $G(\bar{X}_b)$ é maior que zero e sendo assim seleciona-se X_a como nó de decisão. Em outras palavras, o atributo X_a é escolhido para constituir a árvore (HULTEN *et al.*, 2001).

Sendo assim, é possível utilizar o HT como classificador incremental que utiliza uma estrutura de árvore de decisão adaptada para lidar com fluxos de dados.

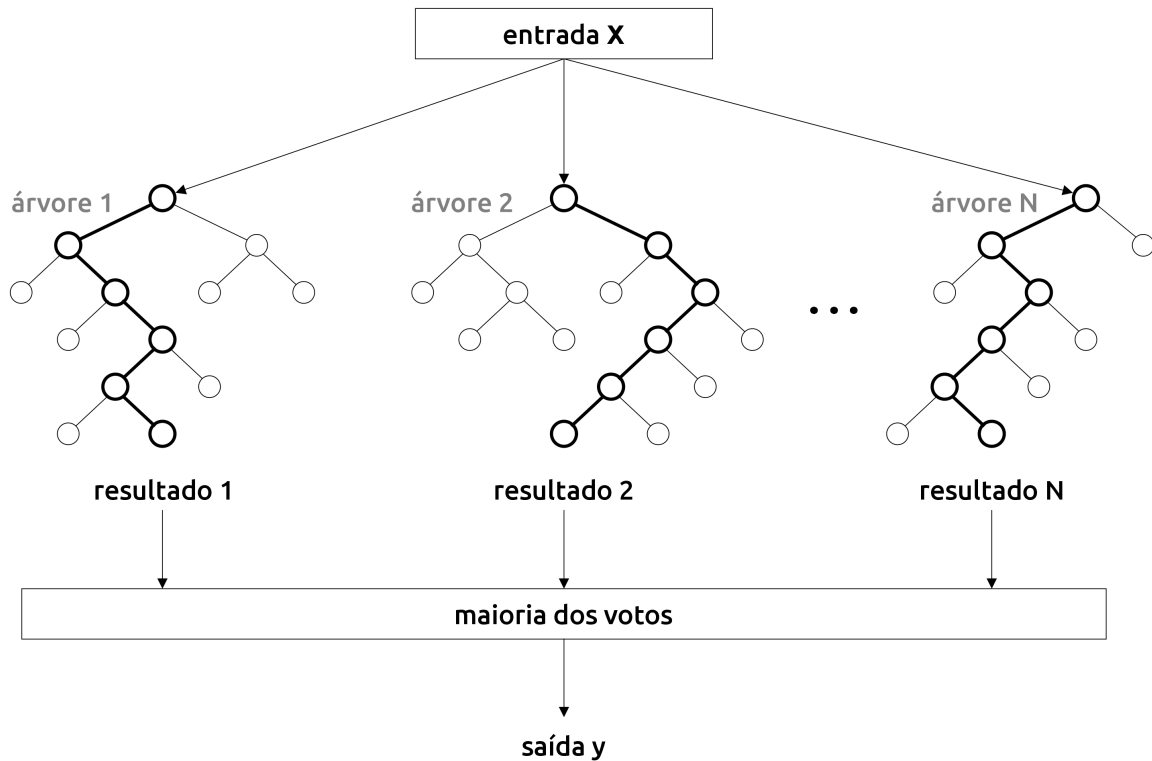
2.3.1.3 Adaptive Random Forest (RF)

Uma floresta aleatória, ou *Random Forest*, é um tipo de classificador *ensemble*. Esse tipo de classificador recebe esse nome por ser um junção, ou aglomerado de múltiplos classificadores e algoritmos mais simples que trabalham juntos, em paralelo, como um classificador mais robusto. Como o nome sugere, o *random forest* se trata de um conjunto de múltiplas árvores de decisão que assim formam uma "floresta". Cada árvore classifica a mesma nova amostra e o resultado desta classificação conta como um voto para determinada classe. Com um número suficientemente grande de árvores na estrutura do *random forest*, cada árvore vota para uma classe e a classe mais popular é então selecionada como a correta para determinada entrada no classificador. A Figura 4 ilustra como é a visão geral do funcionamento de uma *random forest*.

Para a k -ésima árvore de decisão, um vetor aleatório Θ_k é gerado, independente dos vetores anteriores $\Theta_1, \dots, \Theta_{k-1}$ mas com a mesma distribuição. Uma árvore é gerada usando o conjunto de dados de treino e Θ_k , resultando em um classificador $h(\mathbf{x}, \Theta_k)$ onde \mathbf{x} é um vetor de entrada. Desta forma, uma *random forest* pode ser definida como uma coleção de classificadores estruturados como árvores de decisão $\{h(\mathbf{x}, \Theta_k), k = 1, \dots, N\}$, onde cada árvore gera um voto unitário para uma determinada classe e a classe mais popular é escolhida como saída \mathbf{y} dada uma entrada \mathbf{x} (BREIMAN, 2001).

Contudo, esta abordagem convencional de *random forest* funciona para o caso *in batch* ou *offline*. Logo é necessário uma adaptação do classificador para lidar com fluxo de

Figura 4 – Ilustração do funcionamento de uma floresta aleatória.



Fonte: elaborado pelo autor (2023).

dados, ou seja, trazer o classificador para a perspectiva da classificação incremental. É daí que surge a *Adaptive Random Forest* (RF) ou florestas aleatórias adaptativas. Assim como na seção anterior as árvores de decisão convencionais são transformadas em *Hoeffding Trees* para a classificação incremental, a *Adaptive Random Forest* utiliza múltiplas *Hoeffding Trees* como aprendizes básicos do classificador aliados a operadores de mudança de conceito (*concept drift*) (GOMES *et al.*, 2017). Operadores de *concept drifts* são algoritmos que monitoram o padrão dos dados e mudanças nesse padrão sejam súbitas ou de longo prazo, que fazem parte da natureza do fluxo de dados (e.g. padrão de temperatura muda entre dia e noite), mas que podem gerar classificações equivocadas por parte dos classificadores. A RF se baseia em um esquema de alertas e mudanças de conceitos para cada árvore. Assim que um alerta de mudança de uma árvore é emitido, uma nova árvore de *background* começa a ser contruída. Se a mudança acontecer de fato, essa árvore de *background* substitui a árvore cuja mudança de conceito foi detectada. De tal maneira que as florestas vão se adaptando as mudanças de conceito nos dados.

Desta forma, é possível utilizar o *Adaptive Random Forest* como classificador incremental que utiliza múltiplas árvores de decisão do tipo *Hoeffding Tree* sendo adaptada assim para lidar com fluxo de dados.

2.3.2 *Classificadores (in batch)*

Neste trabalho, os classificadores *in batch* foram utilizados no experimento com abelhas sem ferrão, da espécie jataí (*Tetragonisca fiebrigi*). Como o experimento resultou em classes balanceadas de seis colmeias, onde metade tinha rainha presente e a outra metade não, a classificação se mostra uma forma de criar uma generalização eficiente para separação das classes do problema. Através da classificação *in batch*, no experimento em questão, é possível identificar entre múltiplas colmeias quais que estão sem a presença de rainha e conseqüentemente alertar os apicultores responsáveis para auxiliar numa tomada de decisão eficiente.

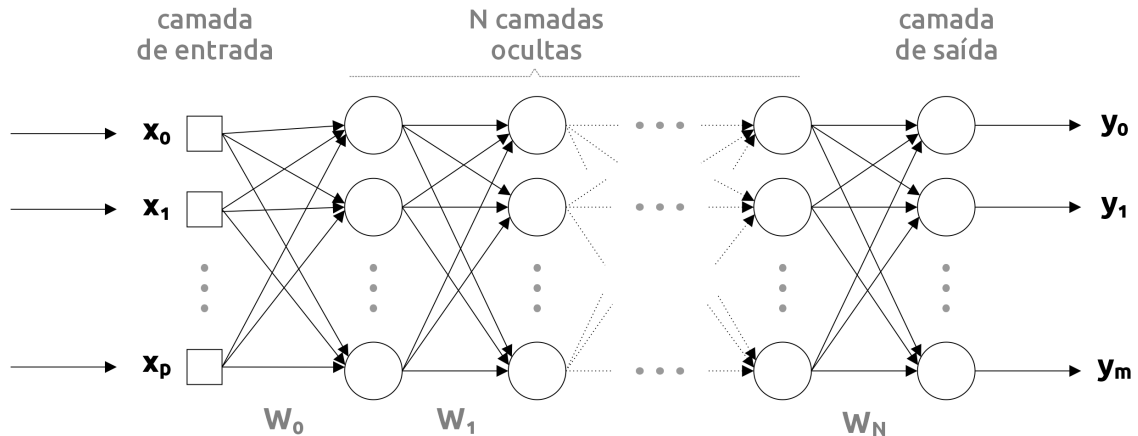
2.3.2.1 *Multilayer perceptron (MLP)*

Perceptron Multicamadas ou *Multilayer Perceptron* (MLP) é um tipo de rede neural que é obtida através da rede básica Perceptron Simples (ROSENBLATT, 1958). O Perceptron Simples é a rede neural mais básica possível com entradas ponderadas através de um função de ativação que gera uma saída. É muito utilizado para problemas linearmente separáveis. Entretanto, para o caso de problemas com separação não-lineares, o perceptron simples não consegue desempenhar bem por não possuir camadas internas ou camadas ocultas em sua arquitetura. Daí nasce o MLP. O MLP possui N camadas ocultas, cada uma com seu próprio número de neurônios e pesos sinápticos associados. Além disso utiliza a estratégia de retro-propagação para minimizar o erro do aprendizado. Sendo assim, o MLP é capaz de generalizar bem uma infinidade de problemas até mesmo os mais complexos. Além de permitir uma gama de diversas arquiteturas com uma variedade de camadas ocultas e número de neurônios em suas camadas.

Por esses motivos, o MLP é um dos modelos de aplicação de redes neurais mais populares, sendo a mais popular no caso de modelos que utilizam retro-propagação como algoritmo de treinamento (RAMCHOUN *et al.*, 2016). A Figura 5 mostra uma arquitetura geral de um MLP. Onde $\mathbf{X} = (x_0, x_1, \dots, x_p)$ é a camada de entrada com p neurônios, \mathbf{W}_i , $i = 0, 1, \dots, N$ são os pesos sinápticos entre as camadas $i - 1$ e i e $\mathbf{Y} = (y_0, y_1, \dots, y_m)$ a camada de saída com m neurônios. Em termos mais práticos, a rede recebe uma entrada \mathbf{x} com p atributos e após processar através das N camadas ocultas, gera uma saída \mathbf{y} com m classes possíveis. A classe escolhida é representada pela saída com o maior valor de saída e que seja superior aos demais dada pela Equação 2.7.

$$\hat{y} = \arg \max_{0 \leq j \leq m} y_j \quad (2.7)$$

Figura 5 – Ilustração da arquitetura básica de um MLP.



Fonte: elaborado pelo autor (2023).

Os pesos da rede são inicializados com valores aleatoriamente pequenos. O algoritmo de retro-propagação calcula o gradiente local do erro e muda os pesos em direção ao vale do gradiente local. Assim, espera-se que os pesos possam convergir para um mínimo global do erro, dado uma quantidade suficiente de dados de treino (GARDNER; DORLING, 1998). O passo-a-passo do funcionamento de um MLP é resumido a seguir.

1. inicializar os pesos da rede de forma aleatória,
2. apresentar a rede o primeiro vetor de entrada através dos dados de treino,
3. progagar o vetor de entrada através da rede para obter a saída,
4. calcular o erro comparando a saída da rede com o valor desejado (*target*),
5. retro-propagar o sinal através da rede,
6. ajustar os pesos para minimizar o erro,
7. repetir os passos 2-7 com o próximo vetor de entrada até que o erro seja suficientemente pequeno ou até o fim dos dados de treino.

Redes neurais do tipo MLP podem assumir diversas arquitetura e conseqüentemente se adequar a diversos tipos de problema. Uma arquitetura 10:256:128:2, por exemplo, se trata de um MLP com dez neurônios na camada de entrada, duas camadas ocultas com 256 e 128 neurônios (respectivamente) e dois neurônios na camada de saída. Ou seja, isso indica que o

problema em questão lida com uma entrada com 10 atributos ou características (*features*) e se trata de uma classificação binária uma vez que a camada de saída só possui dois neurônios.

2.3.2.2 *Extreme Learning Machine (ELM)*

Uma máquina de aprendizado extremo ou *Extreme Learning Machine* (ELM) (HUANG *et al.*, 2004) é uma rede neural não-recorrente (*feedforward*), caracterizado por ter apenas uma camada oculta (*hidden layer*) em sua arquitetura. Em outras palavras, sua arquitetura é de uma rede neural não-recorrente com única camada oculta ou *Single-hidden Layer Feedforward Neural Network* (SLFN).

Máquinas de Aprendizado Extremo foram originalmente desenvolvidas justamente para o problema de lentidão no aprendizado de algoritmos baseados em gradiente para melhora iterativa dos parâmetros da rede. Ele configura todos os parâmetros dos neurônios ocultos de maneira randomica e determina os pesos de saída de forma analítica (LENG *et al.*, 2015). A camada de saída (*output layer*) é, então, um sistema linear, onde os pesos das conexões entre a camada oculta (*hidden layer*) e a camada de saída (*output layer*) são configurados computando a matriz pseudo-inversa de Moore-Penrose (Zhu *et al.*, 2014). A rede neural do ELM tem um aprendizado extremamente rápido devido a sua estrutura simples. Além disso, o caráter randômico deixa com que o ELM não precise necessariamente ajustar os parâmetros de sua rede de maneira iterativa (Zhu *et al.*, 2014).

O ELM é uma SLFN muito eficiente em diversas aplicações e variações e pode ser utilizada para classificação, regressão, *One-class classification* ou classificação de uma classe, processamento de imagens, etc. (DING *et al.*, 2014). Uma breve descrição do funcionamento básico de uma ELM é revisado a seguir.

Dados N amostras de treino distintas $(x_i, t_i) \in R^n \times R^m (i = 1, 2, \dots, N)$, a saída de uma SLFN convencional com \tilde{N} nós ocultos pode ser representada pela equação 2.8.

$$o_j = \sum_{i=1}^{\tilde{N}} \beta_i f(x_j; a_i, b_i), \quad j = 1, \dots, N \quad (2.8)$$

Onde o_j é o vetor de saída da SLFN com respeito a entrada x_j . a_i e b_i são os parâmetros de aprendizado gerados randomicamente para o j -ésimo nó oculto. β_i são os pesos sinápticos de saída ligando o j -ésimo nó oculto com os nós de saída. E $f(x_j; a_i, b_i)$ é a função de ativação da rede neural.

Definindo $a_i \cdot x_j$ o produto interno entre a_i e x_j , a equação pode ser escrita basicamente como mostrado na Equação 2.9. Onde \mathbf{H} é a matriz de saída da camada oculta.

$$\mathbf{H}\beta = \mathbf{O} \quad (2.9)$$

Onde \mathbf{H} , β e \mathbf{O} são definidos como mostrado em 2.10

$$\mathbf{H} = \begin{bmatrix} f(a_1 \cdot x_1 + b_1) & \cdots & f(a_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ f(a_1 \cdot x_N + b_1) & \cdots & f(a_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}}, \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m}, \mathbf{O} = \begin{bmatrix} o_1^T \\ \vdots \\ o_N^T \end{bmatrix}_{N \times m} \quad (2.10)$$

Para minimizar o erro de saída $\|\mathbf{O} - \mathbf{T}\|$, já que os parâmetros de aprendizado são configurados de forma randômica, ou seja, sem considerar os dados de entrada, os pesos sinápticos de saída são determinados de forma analítica encontrando uma solução dos mínimos quadrados, como mostrado em 2.11

$$\beta = \mathbf{H}^\dagger \mathbf{T} \quad (2.11)$$

Neste caso, \mathbf{H}^\dagger é a matriz pseudo-inversa de Moore Penrose da matriz \mathbf{H} . Sendo assim, a saída da rede neural é computada através de transformações matemáticas, evitando muito tempo de processamento com parâmetros ajustados iterativamente (DING *et al.*, 2014) e consequentemente tornando-a mais eficaz e eficiente principalmente em tempo computacional.

2.3.2.3 AdaBoost (ADB)

O *AdaBoost* (ADB) é um algoritmo de classificação do tipo *Boosting*. Esse tipo de algoritmo se refere a um problema de geral de produzir uma regra de predição bastante acurada baseada em uma combinação de múltiplas regras gerais ou princípios básicos que não são tão acurados assim (FREUND; SCHAPIRE, 1997). Entretanto, o *Boosting* possui dois problemas principais: como ajustar o conjunto de treino para permitir os classificadores fracos realizem o treinamento adequado com eles e como combinar os classificadores fracos em um classificador forte. Para resolver estes problemas que surgiu o algoritmo *AdaBoost* (CHENGSHENG *et al.*, 2017).

O algoritmo do ADB diz respeito a um procedimento iterativo que tenta aproximar um classificador bayesiano $C^*(x)$ através da combinação de múltiplos classificadores "fracos". Iniciando com as amostras de treino não ponderadas, o ADB produz um classificador, que pode ser uma árvore de decisão, por exemplo, de forma que este classificador seja capaz de retornar os rótulos das classes. Se um ponto amostral de treino for classificado de forma errada, o peso sobre aquele ponto amostral é aumentado (*boosted*). Neste caso, um segundo classificador é então produzido com os novos pesos. E se repete este mesmo procedimento caso hajam novos erros de classificação durante o treinamento. O número de classificadores pode chegar até 1000, cada um deles recebendo uma pontuação de forma que o classificador final é formado por uma combinação linear dos classificadores de cada estágio (HASTIE *et al.*, 2009).

2.4 Outros conceitos

Nesta seção são abordados especificamente outros métodos que não fazem parte direta da análise de desempenho central deste trabalho mas que foram de alguma forma utilizados. Especificamente, o algoritmo k-médias na Seção 2.4.1 que foi usado para uma análise de agrupamentos que foi realizada nos dados do experimento com abelhas sem ferrão. Além dos índices de validação dos agrupamentos gerados pela aplicação do algoritmo k-médias na Seção 2.4.2. No caso deste trabalho, deseja-se justamente encontrar padrões ocultos que possam ser determinantes nos dados de abelhas sem ferrão. Se espera que as classes (com e sem rainha) representem padrões de divisão entre agrupamentos mais significativos do que padrões individuais nos dados de cada uma das seis colônias.

2.4.1 *k*-médias

O k-médias ou *k-means* é um algoritmo de aprendizado não-supervisionado muito popular para o problema de *clusterização* (SINAGA; YANG, 2020). *Clusterização* diz respeito a uma maneira de classificar dados crus razoavelmente e buscar por padrões ocultos que podem existir em conjuntos de dados. Processando os dados de forma que amostras semelhantes fiquem agrupadas em um mesmo aglomerado de dados e amostras diferentes em aglomerados de dados diferentes (NA *et al.*, 2010).

O objetivo do k-médias é dividir um conjunto de dados de tamanho N em k agrupamentos, com $k \ll N$. Para isso, o algoritmo conta com o auxílio de centroides inicialmente

dispostos aleatoriamente no espaço dos dados. Cada agrupamento é definido por um critério de similaridade (e.g. distância euclidiana) e os dados são então divididos nos k agrupamentos. Em seguida os centróides são atualizados como o ponto médio das amostras de cada agrupamento. Então as amostras pertencentes a cada agrupamento são atualizadas em novos conjuntos das amostras mais próximas aos novos centróides. Continuando assim iterativamente até que a convergência seja alcançada, isto é, quando os centróides não se atualizarem mais entre uma iteração e outra.

Formalmente (BRAGA, 2020), temos que dado um conjunto W de k centróides, onde p é a dimensão do vetor de dados como mostrado na Equação 2.12.

$$W = \{w_i\}_{i=1}^k \mid w_i \in \mathbb{R}^p \quad (2.12)$$

O agrupamento V_i associado ao centróide w_i é definido pela Equação 2.13, onde x é um vetor de atributos e $\|x - w_i\|$ indica a distância euclidiana entre o vetor de atributos x e o i -ésimo centróide w_i .

$$V_i = \{x \in \mathbb{R}^p \mid \|x - w_i\| < \|x - w_j\|, \forall j \neq i\} \quad (2.13)$$

Inicialmente selecionando k vetores aleatórios no espaço dos dados como centróides. O grupo V_i de cada centróide w_i é então determinado. O algoritmo então calcula a nova posição de cada protótipo w_i como o ponto médio das amostras no grupo V_i , conforme ilustrado na Equação 2.14.

$$w_i = \frac{1}{N_i} \sum_{x \in V_i} x \quad (2.14)$$

Então os grupos V_i e centróides w_i , $i = 2, \dots, k$ são recalculados repetidamente até a convergência do algoritmo. A convergência ocorre quando a posição dos centróides w_i não é atualizada entre uma iteração e outra ou quando um número máximo de iterações é atingido.

2.4.2 Índices de validação dos agrupamentos

Para avaliação dos agrupamentos, foram utilizados três índices diferentes. A pontuação *Silhouette* (ROUSSEEUW, 1987), o índice de Calinski-Harabasz (CALIŃSKI; HARABASZ,

1974) e o índice de Davies-Bouldin (DAVIES; BOULDIN, 1979). Os valores dos índices são calculados para vários números de agrupamentos k e aquele que retorna os melhores valores em cada caso, é o número indicado de agrupamentos. Neste trabalho utilizamos os três para uma maior convicção do número correto de agrupamentos em análise.

A primeira das métricas, a pontuação *Silhouette* é um método de validação da consistência de agrupamentos que varia entre -1 e 1, sendo que quanto mais próximo de 1, melhor o resultado. Se o resultado é 1, isso significa que os agrupamentos estão bem separados e se distinguem claramente. Se o resultado for zero, então a distância entre os agrupamentos não é significativa. Enfim, se o resultado for -1 geralmente indica um erro, isto é, uma amostra associada a um agrupamento errado. O cálculo dessa pontuação é relativamente simples como mostrado na Equação 2.15. Onde a é média da distância intra-agrupamento e b é a média da distância entre-agrupamentos.

$$Silhouette = \frac{(b - a)}{\max\{a, b\}} \quad (2.15)$$

Dando sequência, temos o índice de Calinski-Harabasz (CH). Essa medida é definida como a taxa da soma da dispersão entre agrupamentos e a dispersão intra agrupamentos. Para um conjunto de dados de tamanho N com k número de agrupamentos, o índice CH é calculado conforme a Equação 2.16. Onde B_k é a matriz de dispersão entre agrupamentos e W_k a matriz de dispersão intra agrupamentos. O operador $tr(\cdot)$ calcula o traço da matriz. O índice retorna valores maiores conforme os agrupamentos são mais densos e mais bem-separados.

$$CH(k) = \frac{tr(B_k)/(k - 1)}{tr(W_k)/(N - k)} \quad (2.16)$$

Por último, o índice de Davies-Bouldin (DB). Este índice é definido como a medida de similaridade média de cada agrupamento com o agrupamento mais semelhante a ele, onde a similaridade é medida pela taxa das distâncias entre e intra agrupamentos. Portanto, agrupamentos menos dispersos e mais bem-separados resultam em melhores valores. Neste caso, o valor mínimo é zero, com menores valores indicando melhores resultados no número de agrupamentos. O índice de Davies-Bouldin é calculado conforme a Equação 2.17.

$$DB = \frac{1}{N} \sum_{i=1}^N \max_{j \neq i} \frac{S_i + S_j}{M_{i,j}} \quad (2.17)$$

Onde $M_{i,j}$ é a separação entre o i -ésimo e o j -ésimo agrupamento e S_i é a dispersão do agrupamento i . Vemos que, idealmente, o valor de $M_{i,j}$ deve ser o maior possível, uma vez que deseja-se que os agrupamentos sejam bem separados entre si. Enquanto isso, o valor de S_i deve ser o menor possível, uma vez que deseja-se que cada agrupamento seja o menos disperso possível. Por essa razão que, como vemos na Equação 2.17, quanto menor o valor de DB, melhor o resultado para aquela quantidade de agrupamentos.

2.5 Sumário do capítulo

Neste capítulo, tratamos de conceitos importantes diretamente relacionados a este trabalho. Dentre eles, destacamos a Apicultura de Precisão (Seção 2.1) e sua importância para o bem-estar de colmeias de abelhas. Além dos conceitos e trabalhos relacionados com Bioacústica (Seção 2.2), padrões de sinais que transmitem informações importantes e valiosas e que podem ser detectadas de forma não-invasiva. Ademais, foram abordados os classificadores utilizados (Seção 2.3), divididos entre classificadores incrementais (Seção 2.3.1) e classificadores *in batch* ou convencionais (Seção 2.3.2). Por fim, na Seção 2.4, foram apresentados outros conceitos e métodos que foram utilizados nesta dissertação embora não façam parte da análise de desempenho central deste estudo.

3 MATERIAL E MÉTODOS

Este capítulo trata dos materiais utilizados e da metodologia empregada neste estudo. Primeiramente, na Seção 3.1 é falado sobre o objeto de estudo desta dissertação. Em seguida é especificado o delineamento do trabalho (Seção 3.2), que é subdividido de acordo com cada experimento: com abelhas melíferas (Seção 3.2.1) e com abelhas jataí (Seção 3.2.2). Dando continuidade, na Seção 3.3 é discutido sobre os detalhes da avaliação de desempenho dos classificadores estudados. E por último, na Seção 3.4, estão detalhados quais os testes estatísticos utilizados para comparar os resultados de cada classificador.

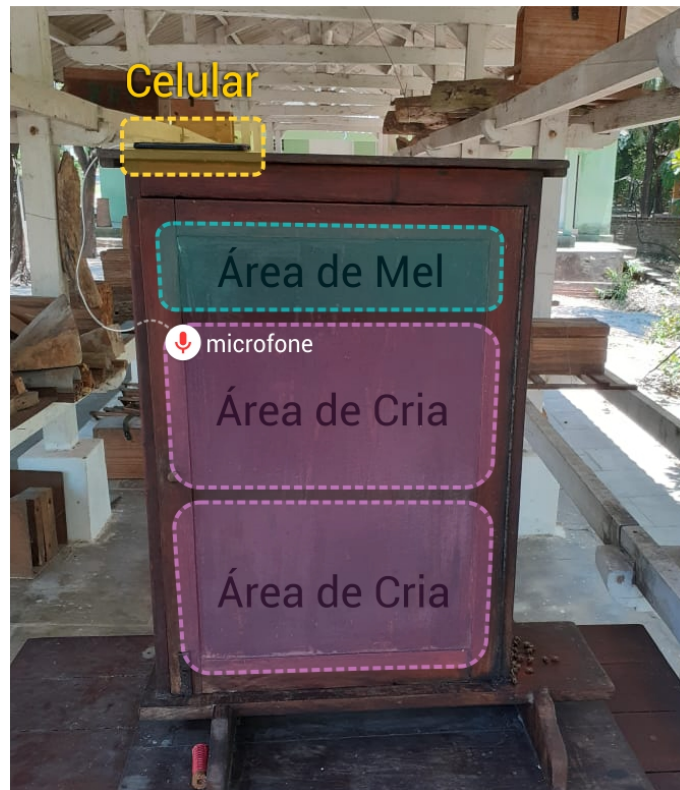
3.1 Objeto de estudo

Sobre o objeto de estudo, neste trabalho trata-se, de forma geral, de espécies de abelhas com e sem ferrão comuns no ecossistema brasileiro.

Inicialmente com uma colmeia de abelhas melíferas (com ferrão), geramos um conjunto de dados de áudios de uma colônia do tipo africanizada (*Apis mellifera*) em uma colmeia de observação do Setor de Abelhas da UFC, localizado no Campus do Pici em Fortaleza-CE (Figura 6). O áudio da referida colmeia foi gravado durante seis dias sempre no período da tarde, onde nos cinco primeiros dias de gravação, a colmeia estava com rainha presente e com o funcionamento normal de suas atividades. No sexto dia a rainha foi retirada da colmeia pela manhã para simular uma perda súbita da rainha em uma colmeia saudável. A remoção foi feita durante a manhã para que quaisquer interferências humanas na colônia fossem dissipadas até o período da tarde, quando a gravação foi feita. A coleta dos dados e a validação de presença da rainha e do bem-estar da colônia, além do manuseio da colmeia, foram todos realizados com auxílio de um apicultor local. Ao todo, foram coletadas mais de 23h de gravação.

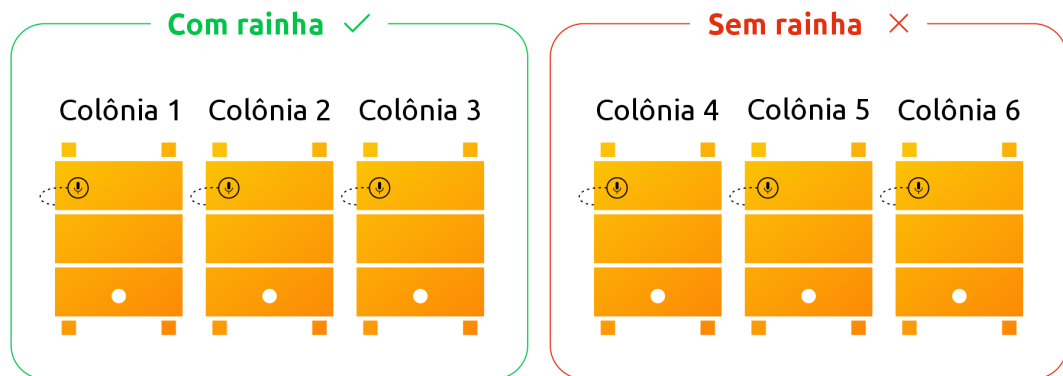
Em seguida, para as abelhas sem ferrão, foram obtidos conjuntos de dados de seis colmeias de abelhas jataí (*Tetragonisca fiebrigi*) localizadas em um meliponário da *start up* mais abelhas, situado em Viamão - RS. Os dados foram disponibilizados por Alex Otesbelgue, que realizou as gravações para o seu trabalho de conclusão de curso em ciências biológicas. Os áudios foram gravados durante cerca de 15 minutos por dia durante 4 dias para cada uma das 6 colmeias analisadas. Esta experimentação tem uma abordagem diferente onde metade (3) das colônias tinham rainha e a outra metade não (Figura 7). Tratando-se assim de uma abordagem de monitoramento de entre colônias. Totalizando cerca de 6h de gravação.

Figura 6 – Colmeia de observação para captura dos áudios de abelhas melíferas.



Fonte: elaborado pelo autor (2023).

Figura 7 – Esquema para captura dos áudios de abelhas sem ferrão.



Fonte: elaborado pelo autor (2023).

Desta forma, foram obtidas amostras de áudio constituindo um delineamento entre a classe positiva (1 - rainha presente) e negativa (0 - rainha ausente) para ambos os experimentos descritos. A diferença, com as abelhas melíferas, é que o sistema possui a abordagem sob o ponto de vista de uma única colmeia monitorada que subitamente perdeu sua rainha. Já com as abelhas jataí sem ferrão, a abordagem é de múltiplas colônias onde algumas possuem rainha presente e outras não. Em ambos os casos foram utilizados microfones simples de fones de

ouvido convencionais e um gravador padrão de *smartphones*.

3.2 Delineamento

O sistema sob estudo pode ser definido, de maneira geral, como um classificador de presença de rainha. O sistema tem como entrada dados de áudio gravados dentro da colmeia. No caso das abelhas melíferas, a entrada são em forma de fluxos de dados. Enquanto isso, no caso das abelhas jataí, a entrada são lotes de dados. Na saída, o sistema retorna as métricas de avaliação do classificador (e.g. acurácia, precisão). Expandindo um pouco mais a perspectiva, o sistema é subdividido em três etapas: amostragem, extração de *features* e a classificação propriamente dita.

A etapa de amostragem está relacionada aos fatores do estudo. O número de janelas e tamanho de janela definem a dimensão dos dados de áudio que são transmitidos e processados pela rede diariamente e, conseqüentemente, o tamanho do conjunto de dados total. A seleção de cada janela no áudio original é feita de forma randomizada. Ou seja, as janelas são escolhidas aleatoriamente do arquivo de áudio original. No caso dos classificadores incrementais, a etapa de amostragem entra na análise de desempenho através de variações na combinação dos fatores. Já no caso dos classificadores *in batch*, a etapa de amostragem é fixa em 200 janelas de 1 segundo, e é apenas uma forma de reduzir e padronizar a dimensão total dos áudios que serão passados para os classificadores.

Na segunda etapa, utilizou-se a biblioteca *pyAudioAnalysis* (GIANNAKOPOULOS, 2015) para a extração de *features* do áudio. A biblioteca é *open-source* e feita para linguagem *python*. Permitindo extrair até 68 *features* do áudio. Dentre as características extraídas existem oito métricas do sinal (e.g. energia, entropia, centroide espectral), que são concatenados com os *Mel Frequency Cepstral Coefficients* ou MFCCs. Estes últimos, para extração de características do áudio, representam a técnica mais popular na literatura. Esta técnica transforma o sinal de áudio em séries compactas de parâmetros que representam o sinal original e é baseada na própria forma do ouvido humano de processar os sinais recebidos. Para a extração de características do áudio original foram usados *frames* de 50 ms com sobreposição de 25 ms. Desta forma, cada segundo amostrado do áudio original produz 39 pontos amostrais com os valores calculados de cada uma das características.

Na terceira e última etapa, temos os classificadores propriamente. Para cada experimento foram escolhidos três classificadores distintos. Todavia, como cada experimento possui seu próprio conjunto de classificadores, eles serão descritos em mais detalhes nas Seções 3.2.1 e

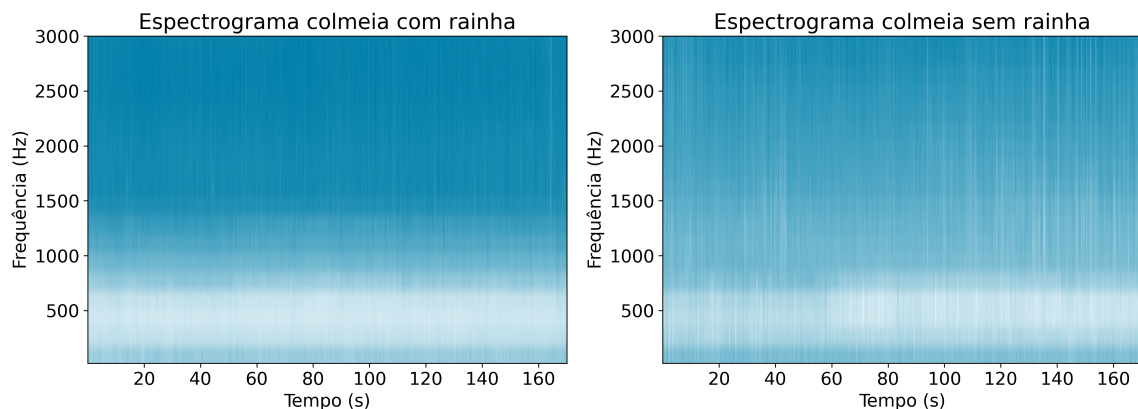
3.2.2 respectivas a cada experimento.

Todos os experimentos foram executados em um computador de processador *Intel core i5* com 8GB RAM.

3.2.1 *Abelhas melíferas*

Para o experimento com abelhas melíferas, temos apenas uma colmeia que com o passar do tempo e do monitoramento, perde sua rainha. O espectrograma na Figura 8 compara o comportamento do áudio no tempo e na frequência durante o último dia de gravação em que a rainha estava presente e o dia em que a rainha foi removida da colônia. É possível notar, ainda na Figura 8, que no caso com rainha a energia é mais concentrada entre 20 e 600 Hz, aproximadamente, o que é bem próximo das faixas de frequência de comportamento padrão de uma colônia. Entretanto, no espectrograma do dia sem rainha, é possível ver rajadas de energia em forma de linhas verticais nas faixas de frequência acima de 500 Hz, além de uma concentração um pouco menor e menos estável nas faixas de 20 a 600 Hz se comparado com o dia em que a colônia tinha rainha presente. A diferença visual é clara entre os dois casos.

Figura 8 – Espectrograma de um intervalo de alguns segundos do último dia com rainha (esquerda) e o dia sem rainha (direita).



Fonte: elaborado pelo autor (2023).

Como no caso deste experimento estamos lidando com dados de um monitoramento periódico e temporalmente sequenciado de uma única colônia, ou seja, fluxo de dados, o tipo de abordagem de classificação estudada para este experimento é a classificação incremental (*online*). Esta abordagem tem uma melhor consonância com o tipo de delineamento do experimento de abelhas melíferas que utiliza fluxo de dados. O classificador vai receber os dados amostra por amostra, avaliando, predizendo e se adaptando em tempo de execução. Após percorrer todo o

conjunto de dados, o classificador retorna as métricas de avaliação. Por conta da natureza dos dados e desbalanceamento das classes a classificação binária convencional (*in batch* ou *offline*) não é a mais adequada. Sendo assim, foi definida a adoção da classificação incremental que se mostra mais alinhada com o experimento. Graças a isso, a análise de desempenho nesse caso pode ser mais profunda, envolvendo não apenas o classificador usado no sistema mas também a combinação de fatores do estudo.

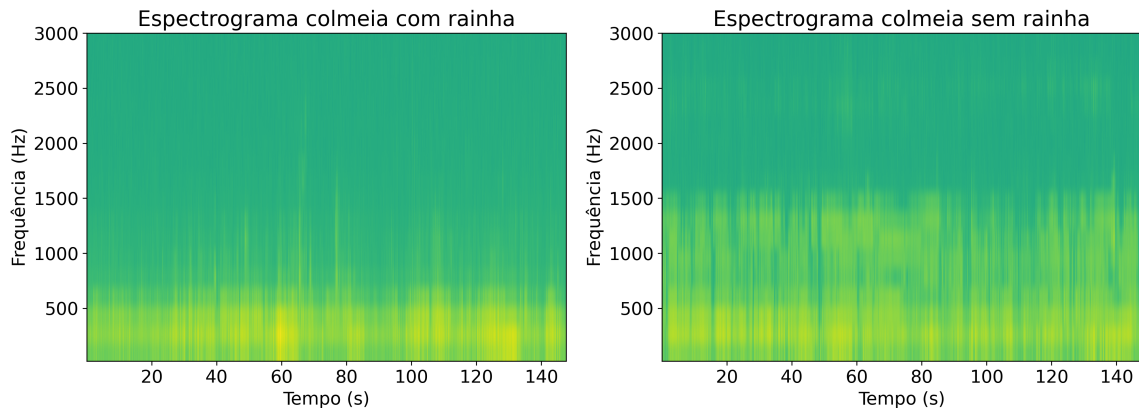
Neste experimento foram analisados três classificadores: *Hoeffding Tree* (HT) - uma árvore de decisão rápida utilizando o limite de *Hoeffding*, *Naive Bayes* (NB) - um classificador baseado no teorema de Bayes e *Adaptive Random Forest* (RF) uma adaptação do *Random Forest* tradicional para abordagem incremental. Todos usando a biblioteca *scikit-multiflow* e com os parâmetros predefinidos pela biblioteca. Para o RF foi utilizado um *random state* igual a 42, para fins de reproducibilidade. Para validação do fluxo de dados com o classificador incremental, utilizamos a metodologia *prequential*. Na abordagem *prequential*, cada instância do fluxo de dados, uma a uma, é usada primeiramente para testar e em seguida atualizar incrementalmente o modelo. A análise de desempenho é feita individualmente e, com os resultados obtidos, é feita uma comparação posterior entre os classificadores.

3.2.2 *Abelhas jataí*

Já com o experimento de abelhas jataí sem ferrão, temos uma perspectiva um pouco diferente da anterior. Neste caso as classes são bem balanceadas onde metade das colônias tinham rainha presente no momento da gravação e a outra metade não. Ou sejam trata-se de uma abordagem entre colônias. O espectrograma na Figura 9 relaciona, quanto ao tempo e a frequência, um intervalo de alguns segundos do quarto dia de gravação de uma das colônias com rainha e uma das colônias sem rainha, escolhidas de forma aleatória. Nota-se que no caso com rainha, há uma maior predominância da energia na faixa entre 0 e 500 Hz, com algumas leves rajadas em frequências superiores no centro horizontal da figura, que muito possivelmente representam *outliers* do sinal. Já no caso da colmeia sem rainha, temos um padrão de energia concentrado nas faixas entre 0 e 1500 Hz, mostrando um padrão não tão estável também. A diferença visualizada é bem clara em relação aos dois casos. Também é interessante comparar o padrão de uma colmeia sem rainha de abelhas sem ferrão (Figura 9) com a ausência de rainha com ferrão (Figura 8). Vemos que o padrão bioacústico de cada espécie é naturalmente distinto, mesmo que existam outros fatores, além da espécie, determinantes para esta diferença e.g. clima,

número de dias sem rainha, força da colônia, quantidade de cria.

Figura 9 – Espectrograma de um intervalo de alguns segundos de uma das colônias com rainha (esquerda) e uma sem rainha (direita).



Fonte: elaborado pelo autor (2023).

No caso deste experimento, tem-se uma abordagem com múltiplas colônias em observação e não mais apenas uma onde aquela mesma perde sua rainha. Por conta disso, a abordagem da classificação incremental não é tão adequada. Em contrapartida, a classificação *in batch* convencional é interessante para este caso, uma vez que os dados estão bem balanceados e se busca encontrar um limite de decisão para separar as classes positivas e negativas para presença de rainha. Infelizmente neste caso, os fatores de amostragem do estudo não tem tanta relevância para análise uma vez que se trata de uma abordagem *offline* e uma vez que os dados possuem um número de amostras suficientemente grande, os modelos podem desempenhar de maneira satisfatória. Por isso, os fatores da amostragem foram fixados em 200 janelas de 1 segundo (200 & 1) para este experimento.

Contudo, sabemos que, como lidamos com colônias diferentes, com assinatura distintas e desempenhando serviços não necessariamente iguais, é preciso considerar se os classificadores estão de fato traçando uma separação entre as classes desejadas ou apenas dividindo cada uma das colônias. Em outras palavras, é preciso garantir que os dois agrupamentos relativos a presença de rainha são mais relevantes do que os seis agrupamentos respectivos a cada colônia. Ou seja, os dados são mais bem separados em dois grupos (um para cada classe com/sem rainha) ou em seis grupos distintos (um para cada colônia, com suas individualidade). Para isso será utilizado uma abordagem não supervisionada usando o algoritmo k-médias para $2 \leq k \leq 6$ e avaliado utilizando três coeficientes de similaridade/dissimilaridade entre e intra *clusters* para validar a natureza da separação dos dados. Os coeficientes utilizados foram o *silhouette* (quanto maior melhor), o Calinski-Harabasz (quanto maior, melhor) e o índice de

Davies-Bouldin (quanto menor, melhor). Para determinar então o número de agrupamentos que os dados melhor são separados, foram analisados os resultados para estes três índices e o número k ideal é aquele que apresentar melhor resultado geral.

Ademais, neste experimento também serão usados três classificadores. Neste caso, três redes neurais: o *Multilayer Perceptron* (MLP) - uma rede neural de alimentação direta baseada no perceptron simples mas com múltiplas camadas ocultas e retro-propagação; o *AdaBoost* (ADB) - um classificador de *boosting* que é formado por combinação linear de múltiplos classificadores simples; e o *Extreme Learning Machine* (ELM) - que é uma rede neural de alimentação direta cuja saída é calculada de forma analítica baseado na matriz pseudo-inversa de Monroe-Penrose. Todos usando a biblioteca *scikit-learn* e com os parâmetros predefinidos pela biblioteca. Além disso, para o MLP foi utilizada uma arquitetura com uma única camada oculta com 100 neurônios. Para o ELM, foi utilizada uma camada oculta com 1000 neurônios. A análise de desempenho é feita individualmente e apenas variando os classificadores (como já foi dito, os fatores de amostragem não variam nesta abordagem), com os resultados obtidos, é feita uma comparação posterior entre os classificadores.

3.3 Avaliação de desempenho dos classificadores

Selecionamos dois fatores que tratam da etapa de amostragem do áudio: (i) número de janelas e (ii) tamanho da janela. Tendo em vista que as gravações de áudio geram arquivos pesados e extensos, é recomendável buscar uma melhor maneira de selecionar intervalos diários para que o áudio seja coletado e transmitido sem a necessidade de enviar extensas horas de gravação através da rede. No caso dos classificadores incrementais, os níveis escolhidos para o fator (i) são 10, 100 e 200 janelas. Para o fator (ii) são 1, 2 e 3 segundos. No caso dos classificadores convencionais *in batch*, os níveis não variam, são fixos em 200 janelas para o fator (i) e 1 segundo para o fator (ii). Neste trabalho será utilizada uma notação para determinada combinação dos dois fatores. No caso, a notação é $(x \& y)$ para representar a combinação de x janelas de y segundos. Por exemplo, $(200 \& 1)$ se refere a 200 janelas de 1 segundo de amostragem de áudio diária.

As métricas da análise são as mesmas para ambos os experimentos: (a) tempo de execução e (b) acurácia do classificador. Além destas, foram computadas outras medidas, como precisão, sensibilidade (*recall*) e a medida $F_{\beta=0,5}$. Muito embora, apenas com (a) e (b) foram realizadas análises estatísticas sistemáticas, ainda assim as outras métricas são importantes

para validação e escolha dos classificadores. Para a classificação, para ambos os experimentos, existem quatro grupos que compõem a matriz de confusão que é usada para determinar as demais métricas de saída dos classificadores. Sendo estes:

- **Verdadeiros Positivos (VP)** - amostras classificadas com a presença de rainha e que de fato a rainha estava presente na colônia;
- **Falsos Positivos (FP)** - amostras que foram classificadas como presença de rainha mas na verdade a rainha estava ausente (Erro do tipo I);
- **Falsos Negativos (FN)** - amostras classificadas como ausência de rainha mas a rainha na verdade estava presente (Erro do tipo II);
- **Verdadeiros Negativos (VN)** - amostras classificadas como ausência de rainha e que de fato estavam sem rainha na colônia.

A acurácia (Equação 3.1) é a taxa de acerto do classificador, calculada basicamente como o número de predições corretas dividido pelo número total de predições realizadas.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

A acurácia é uma das métricas sempre lembradas e bastante utilizadas para avaliar modelos de aprendizado de máquina. Entretanto, em casos de classes desbalanceadas ou com poucos pontos amostrais, os resultados de acurácia podem ser enviesados. Desta forma, é necessário analisar outras métricas para garantir a eficiência do modelo.

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

A precisão (Equação 3.2) é uma métrica muito importante de ser analisada, em especial para este trabalho. Isto porque ela diminui conforme o número de Falsos Positivos cresce. Em outras palavras, ela indica o quanto o modelo está sofrendo erros do tipo I (Falso Positivo). Para o sistema em análise, como estamos lidando com a presença de rainha que é um fator crítico para o bem-estar de um colônia, um erro do tipo I é o mais indesejado. A ausência de rainha passar como um dado normal pode ser muito prejudicial ao sistema. Enquanto é mais "tolerável" (embora também não desejado) os erros do tipo II (Falso Negativo). Uma colônia com rainha presente ser classificada como tendo ausência de rainha, pode gerar um estado de alerta mas não afetaria o sistema. O apicultor responsável se deslocaria até a colmeia e atestaria que foi apenas um alarme falso, o bem-estar da colônia não é afetado criticamente com erros do tipo

II. Com abordagem análoga à precisão, temos o Recall (Equação 3.3). Neste caso, ele diminui conforme a quantidade de falsos negativos cresce. Ou seja, ela indica o quanto o modelo está sofrendo erros do tipo II.

$$Recall = \frac{TP}{TP + FN} \quad (3.3)$$

De maneira geral, a precisão cai conforme existem mais erros do tipo I e o *recall* cai conforme existem mais erros do tipo II.

A medida F_β é a média harmônica ponderada entre precisão e *recall*. A ponderação é feita através do parâmetro β . Se $\beta = 1$, a medida é igual ao tradicional *f-score* (Equação 3.4). Sendo o *f-score* tradicional um balanceamento harmônico entre erros do tipo I e tipo II, logo é uma métrica sensível a qualquer tipo de erro da classificação. O parâmetro β serve para ponderar e tornar a métrica mais sensível a erros que podem ser mais graves que outros, tornando assim uma métrica mais adequada a determinados problemas. Se $\beta \rightarrow 0$ (tende a zero) então a métrica passa a considerar apenas a precisão e se $\beta \rightarrow \infty$ (tende a infinito) a métrica passa a considerar apenas o *recall*.

$$F - Score = 2 \frac{Precision * Recall}{Precision + Recall} \quad (3.4)$$

Conforme já mencionado, para o problema abordado neste trabalho, erros do tipo I são muito mais preocupantes e representam mais riscos ao bem-estar da colônia. Por conseguinte, foi escolhido $\beta = 0,5$ já que a precisão é a métrica mais sensível aos erros do tipo I. E, consequentemente, o $F_{\beta=0,5}$ é uma medida mais sensível a erros do tipo I do que do tipo II, e fornece uma boa perspectiva do quão bom está a classificação do sistema.

3.4 Testes estatísticos

Nesta seção estão descritos brevemente os métodos estatísticos escolhidos para a análise sistemática seja para averiguar se os fatores sob estudo são impactantes nas métricas avaliadas ou para comparar entre os classificadores se possuem diferenças significantes ou não. Os resultados dos experimentos deste estudo não validaram os pressupostos de homogeneidade das variâncias e normalidade dos resíduos em nenhum dos casos. Portanto, não é possível realizar

uma análise de variância (ANOVA) padrão, por se tratar de um método paramétrico que requer a validação destes pressupostos.

Como neste trabalho foi realizada uma análise fatorial tripla (classificador, número e tamanho das janela), o método estatístico escolhido foi a ANOVA ART (WOBBROCK *et al.*, 2011) (AART, neste texto), o qual é mais apropriado para experimentos fatoriais múltiplos não-paramétricos como neste caso. Por se tratar de um teste não-paramétrico, os pressupostos de homogeneidade das variâncias e normalidade dos resíduos não precisam ser validados. Esta análise testa a hipótese nula de que as distribuição testadas advém de um mesmo tipo de distribuição. Se o valor-p resultante for menor que o nível de significância adotado, então podemos rejeitar a hipótese nula e aceitar a hipótese alternativa de que as distribuições dos dados testados são diferentes. O nível de significância adotado para todos os testes estatísticos foi de 0,05 (5%). Os testes foram executados utilizando a linguagem *R*. A AART foi adotada para todos os testes em ambos os experimentos para que não houvessem múltiplos testes diferentes entre os experimentos.

3.5 Sumário do capítulo

Este capítulo apresentou a metodologia e procedimentos realizados no desenvolvimento desta dissertação. Primeiro foi apresentado o objeto de estudo (Seção 3.1). Em seguida, temos o delineamento estratégico deste trabalho (Seção 3.2) que é subdividido para cada uma das espécies estudadas: abelhas melíferas africanizadas (Seção 3.2.1) e abelhas jataí (Seção 3.2.2). Depois, na Seção 3.3, são detalhados os fatores e métricas de avaliação que são consideradas para a análise de desempenho dos classificadores utilizados em cada um dos experimentos. Finalmente, foi apresentado o tipo de teste estatístico utilizado para validação dos resultados e hipóteses (Seção 3.4).

4 RESULTADOS

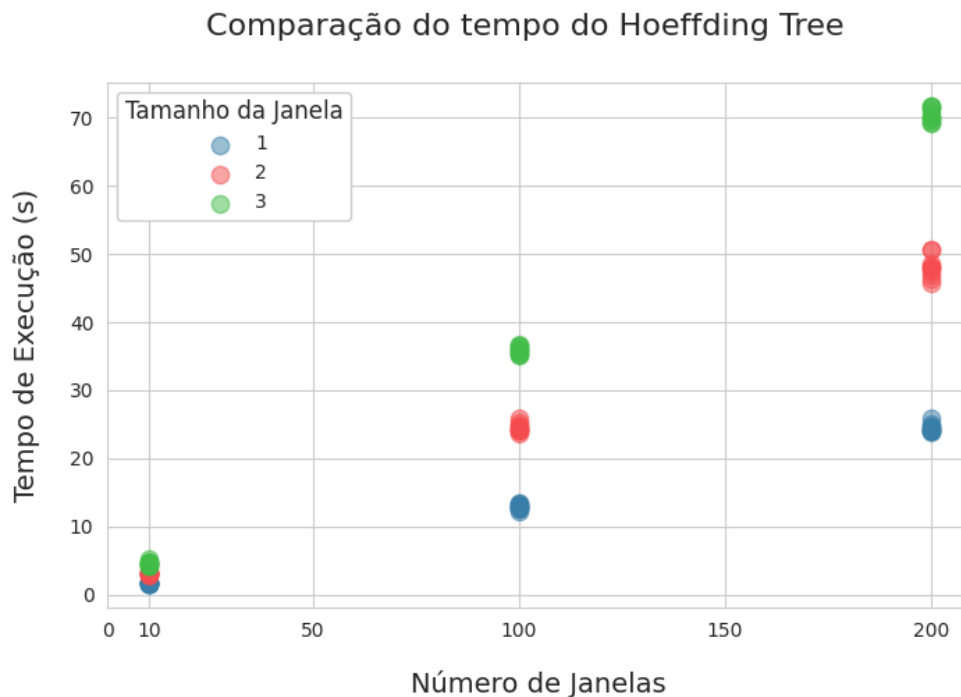
Este capítulo está dividido entre os dois experimentos relativos às duas espécies de abelhas sob estudo. Na Seção 4.1 temos primeiramente os resultados do experimento realizado com abelhas melíferas do tipo africanizadas (abelhas com ferrão) e classificadores incrementais. Na Seção 4.2, trazemos os resultados do experimento com abelhas jataí (abelhas sem ferrão) e classificadores *in batch*. Em seguida, temos uma breve discussão dos resultados obtidos na Seção 4.3.

4.1 Resultados do experimento com a colônia de abelhas com ferrão (*Apis mellifera*)

Para o caso da colônia de abelhas melíferas localizadas em Fortaleza, foram analisados os resultados dos três classificadores incrementais: *Hoeffding Tree* (Seção 4.1.1), *Naive Bayes* (Seção 4.1.2) e o *Adaptive Random Forest* (Seção 4.1.3). Cada subseção trata dos resultados individuais de cada um e ao final (Seção 4.1.4) há uma análise comparativa entre eles.

4.1.1 *Hoeffding Tree*

Figura 10 – Comparativo do tempo de execução do algoritmo *Hoeffding Tree* (HT).



Fonte: elaborado pelo autor (2023).

O tempo de execução do HT (Figura 10) mostrou-se bastante eficiente. O classificador chegou a marcar um pouco mais de 70 segundos de tempo de execução para o caso extremo de 200 janelas de 3 segundos (200 & 3). Além disso é possível observar um comportamento linear ao travar-se um dos fatores (tamanho da janela ou número de janelas) em estudo. O tempo mais eficiente ou menos custoso é observado na combinação de 10 janelas de 1 segundo (10 & 1), onde tem-se um tempo médio de $1,61 \pm 0,11$ segundos.

Tabela 1 – Resultado das métricas (média \pm desvio padrão) de HT.

Combinação	Hoeffding Tree			
	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
(10 & 1)	96,8 \pm 0,7	99,1 \pm 0,5	97,3 \pm 0,7	98,7 \pm 0,4
(10 & 2)	97,2 \pm 0,7	98,9 \pm 0,6	97,8 \pm 0,7	98,7 \pm 0,5
(10 & 3)	97,5 \pm 1,4	98,8 \pm 0,4	98,4 \pm 1,3	98,7 \pm 0,6
(100 & 1)	97,7 \pm 0,3	98,6 \pm 0,3	98,6 \pm 0,2	98,6 \pm 0,2
(100 & 2)	98,1 \pm 0,2	98,9 \pm 0,1	98,8 \pm 0,2	98,9 \pm 0,1
(100 & 3)	98,3 \pm 0,2	99,1 \pm 0,2	98,9 \pm 0,1	99,1 \pm 0,1
(200 & 1)	98,0 \pm 0,1	98,9 \pm 0,1	98,7 \pm 0,1	98,9 \pm 0,1
(200 & 2)	98,3 \pm 0,1	99,1 \pm 0,1	98,9 \pm 0,1	99,1 \pm 0,1
(200 & 3)	98,5 \pm 0,1	99,2 \pm 0,1	99,0 \pm 0,1	99,2 \pm 0,1

Fonte: elaborado pelo autor (2023).

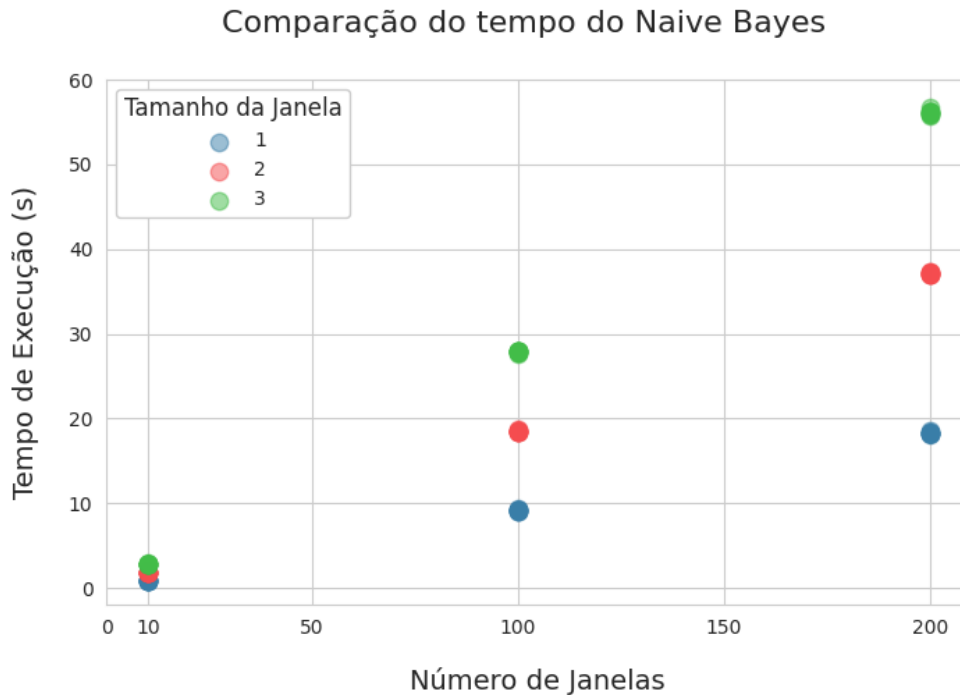
Para a Acurácia (Tabela 1) temos valores sempre acima de 96% para todas as combinações. A acurácia está destacada pois é a métrica escolhida para os testes estatísticos juntamente com o tempo de execução. É possível notar que a acurácia média cresce conforme o nível das combinações de fatores aumentam ao passo que o desvio padrão reduz indicando que as acurácias vão convergindo mais a medida que as combinações de fatores crescem. Ademais, a métrica $F_{\beta=0,5}$ e a precisão se mantiveram sempre acima de 98%, indicando uma baixa ocorrência de erros do tipo I. A AART mostrou que apenas as combinações (100 & 2) e (200 & 1) produzem tempos de execução iguais. Quanto a acurácia, nem todas são iguais estatisticamente. Entretanto, como já mencionado, as acurácias médias são em todos os casos acima de 96% chegando a um máximo de 98,5%. Desta forma, optar por uma combinação que tenha acurácia menor que outra ainda é interessante porque ganha-se bastante tempo em troca de casas decimais de acurácia. Portanto, desta forma é possível ganhar muito tempo de processamento em troca de uma pequena variação na acurácia do modelo, tornando-o menos custoso computacionalmente.

Sendo assim, para o HT, a melhor combinação de fatores seria 10 janelas de 1 segundo (10 & 1) com tempo médio igual a 1,68 segundos. Estatisticamente, esta combinação possui tempo de execução e acurácia menor que as demais combinações. Entretanto, a variação de acurácia média é mínima ($< 2,0\%$), enquanto o ganho de tempo é muito maior (> 60 s).

Ponderando, o ganho em eficiência computacional supera a perda em acurácia em análise.

4.1.2 Naive Bayes

Figura 11 – Comparativo do tempo de execução do algoritmo *Naive Bayes* (NB).



Fonte: elaborado pelo autor (2023).

O tempo de execução do NB (Figura 11) mostra um ótimo desempenho do classificador, não ultrapassando a marca de 60 segundos nem mesmo para o caso extremo de 200 janelas de 3 segundos (200 & 3). Além do mais, a AART mostrou que todas as combinações de fatores geram respostas de tempo de execução diferentes entre si. Assim como para o HT, é possível observar um comportamento linear ao travar-se um dos fatores (tamanho da janela ou número de janelas) em estudo com o NB. Ainda de forma semelhante ao caso anterior, o tempo mais eficiente ou menos custoso é observado na combinação de 10 janelas de 1 segundo (10 & 1), onde tem-se um tempo médio de $0,93 \pm 0,02$ segundos.

Em relação a acurácia do NB (Tabela 2), temos valores muito próximos entre as combinações de fatores com pouca variação média. A menor acurácia média registrada ainda é bastante alta, mantendo a métrica sempre acima de 96%. Outra métrica bastante elevada é a $F_{\beta=0,5}$ que se mantém em torno dos 99% em todos os casos. O mesmo vale para a precisão. Portanto, indicam uma baixíssima incidência de erros do tipo I. Não há nenhuma variação

Tabela 2 – Resultado das métricas (média \pm desvio-padrão) de NB.

Combinação	Naive Bayes			
	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
(10 & 1)	97,0 \pm 1,2	99,7 \pm 0,2	97,0 \pm 1,4	99,0 \pm 0,4
(10 & 2)	96,6 \pm 0,9	99,8 \pm 0,1	96,2 \pm 1,0	99,1 \pm 0,2
(10 & 3)	97,3 \pm 1,7	99,8 \pm 0,4	97,0 \pm 1,7	99,2 \pm 0,7
(100 & 1)	96,9 \pm 0,5	99,7 \pm 0,1	96,5 \pm 0,6	99,0 \pm 0,1
(100 & 2)	96,7 \pm 0,4	99,6 \pm 0,1	96,4 \pm 0,4	99,0 \pm 0,1
(100 & 3)	96,9 \pm 0,4	99,7 \pm 0,1	96,5 \pm 0,5	99,0 \pm 0,1
(200 & 1)	96,6 \pm 0,2	99,7 \pm 0,0	96,2 \pm 0,3	99,0 \pm 0,1
(200 & 2)	96,5 \pm 0,3	99,7 \pm 0,0	96,2 \pm 0,4	99,0 \pm 0,1
(200 & 3)	96,7 \pm 0,2	99,7 \pm 0,1	96,4 \pm 0,2	99,0 \pm 0,1

Fonte: elaborado pelo autor (2023).

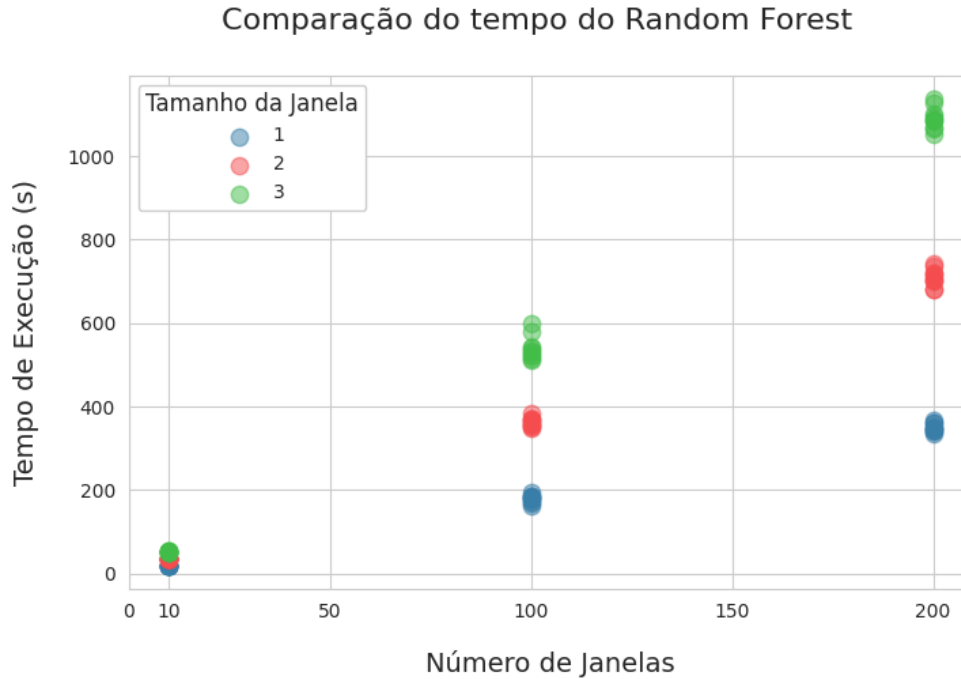
aparente na acurácia média conforme o nível das combinações de fatores aumentam, isto é, não parecem estar correlacionadas de nenhuma forma. Isto é corroborado pelos resultados da AART, que confirmam que as acurácias são estatisticamente iguais entre quaisquer combinações de fatores para o NB.

Consequentemente, utilizando o NB, a melhor combinação de fatores é (10 & 1), sem extra ponderações. Esta combinação possui tempo de execução médio igual a 0,93s. Mostrando ser muito rápido e ainda estatisticamente diferente das demais combinações. Em contrapartida, (10 & 1) produz uma acurácia estatisticamente igual a todas as demais combinações de fatores avaliadas neste estudo. Sendo assim, rearranjar os fatores não causa impacto na métrica de acurácia, mas tem forte impacto na métrica de tempo de execução. Logo, é possível escolher a combinação que corresponda ao menor tempo de execução sem precisar ponderar ou compensar nenhuma variação na acurácia uma vez que ela não depende dos fatores e produz valores estatisticamente iguais para quaisquer combinação destes.

4.1.3 Adaptive Random Forest

O tempo de execução do RF (Figura 12) mostra o alto custo computacional do classificador. O tempo do RF é mais de 10 vezes maior que os outros classificadores avaliados nesse estudo. Este alto custo ocorre devido a complexidade do algoritmo de florestas aleatórias adaptativas que é um método de *ensemble* que são naturalmente mais robustos para problemas mais complexos e que exijam toda esta robustez do algoritmo. Por conta disso, só seria viável aplicar este classificador para o problema proposto utilizando apenas 10 janelas. Conforme os níveis dos fatores aumentam, o tempo escala drasticamente. Ainda assim, mesmo para 10 janelas de 1 segundo (10 & 1), por exemplo, ainda trata-se de um tempo médio de 18,5 \pm 0,5 segundos. Este tempo excede, e muito, o tempo observado para a mesma combinação de fatores

Figura 12 – Comparativo do tempo de execução do algoritmo *Random Forest* (RF).



Fonte: elaborado pelo autor (2023).

nos algoritmos anteriores. O algoritmo gera múltiplas florestas adaptativas, possuindo até mesmo técnicas de mudança de conceito integradas. No entanto, se mostrou demais para o problema em questão que pode ser resolvido com abordagens mais simples como o HT e o NB, por exemplo, e que, conseqüentemente, são mais eficientes em tempo de execução.

Tabela 3 – Resultado das métricas (média \pm desvio-padrão) de RF.

Combinação	Random Forest			
	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
(10 & 1)	96,4 \pm 0,7	97,3 \pm 0,5	98,6 \pm 0,4	97,5 \pm 0,5
(10 & 2)	97,6 \pm 0,5	98,5 \pm 0,4	98,9 \pm 0,3	98,5 \pm 0,3
(10 & 3)	98,0 \pm 0,6	98,5 \pm 0,4	99,2 \pm 0,4	98,6 \pm 0,3
(100 & 1)	98,4 \pm 0,2	99,1 \pm 0,1	99,0 \pm 0,1	99,1 \pm 0,1
(100 & 2)	98,7 \pm 0,2	99,3 \pm 0,1	99,1 \pm 0,1	99,3 \pm 0,1
(100 & 3)	98,8 \pm 0,2	99,4 \pm 0,1	99,2 \pm 0,1	99,4 \pm 0,1
(200 & 1)	98,7 \pm 0,1	99,3 \pm 0,1	99,2 \pm 0,1	99,3 \pm 0,1
(200 & 2)	98,8 \pm 0,1	99,5 \pm 0,1	99,2 \pm 0,1	99,4 \pm 0,1
(200 & 3)	98,9 \pm 0,1	99,5 \pm 0,1	99,2 \pm 0,1	99,4 \pm 0,0

Fonte: elaborado pelo autor (2023).

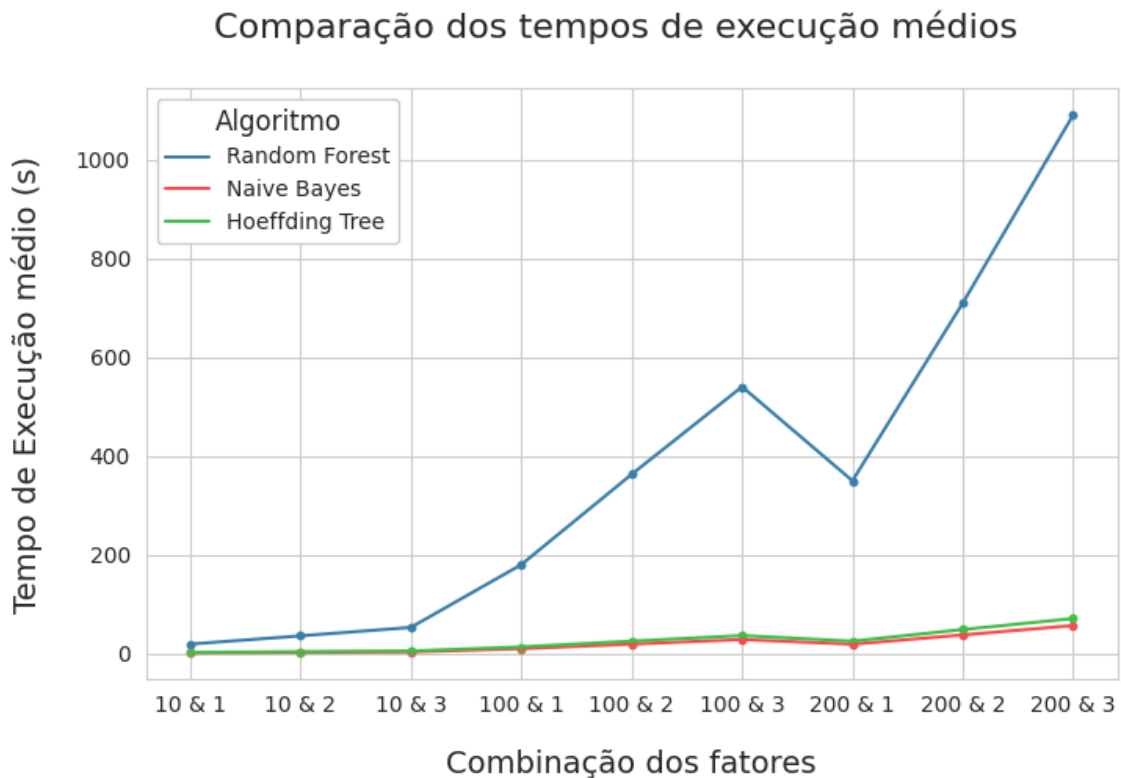
Em relação a acurácia do RF (Tabela 3), temos valores muito altos. Onde também é possível observar um crescimento dos valores de acurácia conforme se escala os níveis dos fatores. Entretanto, para apenas 10 janelas (a única faixa da combinação de fatores onde ainda pode haver uma comparação com os demais algoritmos) os valores de acurácia não se sobressaem dos demais classificadores, tornando o tempo ainda mais um fator limitante. Em outras palavras, o algoritmo

gasta mais tempo para produzir valores de acurácia semelhantes aos demais classificadores. A partir de certa combinação de fatores as acurácias se sobressaem e trazem respostas superiores aos demais. Infelizmente, neste ponto o custo em tempo de execução aumenta de forma que não torna viável a utilização do RF em comparação ao HT ou ao NB.

Desta forma, o RF não se mostra um classificador em condições de concorrer com os demais. É o mais robusto dentre os algoritmos, porém, para o problema em questão, não é eficiente como os anteriores. Utilizando o AART, os resultados mostram que as acurácias do RF são iguais em muitas combinações ao HT, por exemplo. Portanto, o alto tempo de execução não se torna proveitoso de nenhuma forma.

4.1.4 Comparando os 3 classificadores

Figura 13 – Tempos de execução entre os classificadores.

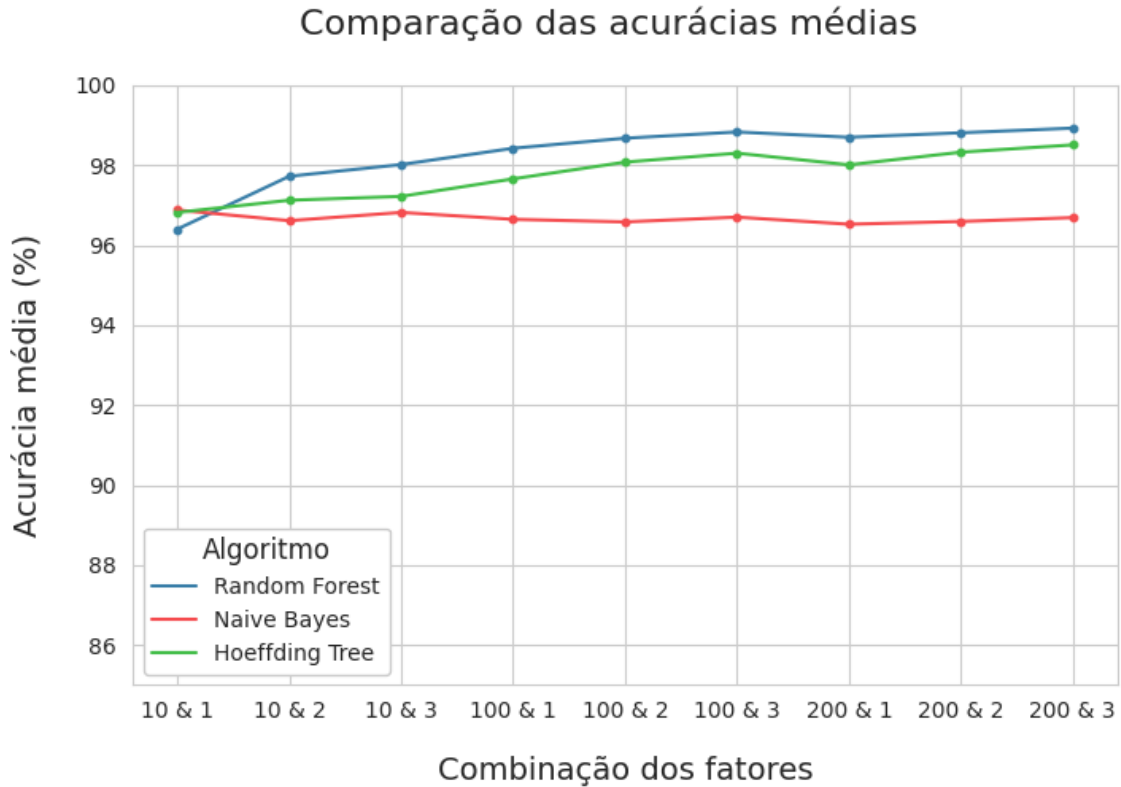


Fonte: elaborado pelo autor (2023).

A Figura 13 ilustra um comparativo dos tempos de execução dos classificadores *Hoeffding Tree* (HT), *Random Forest* (RF) e *Naive Bayes* (NB). Com relação ao tempo de execução, o classificador mais rápido foi o NB seguido bem de perto pelo HT em cada uma das combinações de fatores. De início, o classificador RF desempenha similar aos outros dois mas se

torna muito custoso com o aumento dos níveis dos fatores, chegando ao ponto de ter um tempo de execução cerca de dez vezes maior.

Figura 14 – Acurácia média entre os classificadores.



Fonte: elaborado pelo autor (2023).

Com relação à acurácia (Figura 14), os 3 classificadores apresentam acurácias bem elevadas (acima de 96 %) para todas as combinações de níveis dos fatores e sem grande elevação conforme os níveis dos fatores aumentam. Vale ressaltar que o eixo vertical do gráfico está aproximado, variando de 85 % até 100 % para que seja possível diferenciar as sutis diferenças de acurácia média entre os classificadores. Se fossem apresentadas na escala normal de 0 a 100 %, não seria possível distinguir entre eles dada a faixa de valores apresentadas. As linhas pareceriam linhas retas sobrepostas.

Tanto para o NB, quanto para o HT, a melhor combinação de fatores é (10 & 1). O tempo do NB é 0,93 s e do HT é igual a 1,68 s. As acurácias são respectivamente 96,8 % e 97,0 %. A métrica $F_{\beta=0,5}$ é maior para o NB, indicando a menor ocorrência de erros do tipo I, o que é mais desejável para garantir o bem-estar das abelhas melíferas. A matriz de confusão média do NB e HT para 10 janelas de 1 segundo é mostrada na Figura 15. É possível notar que o NB identifica a ausência de rainha melhor que o HT, enquanto para presença de rainha é o contrário.

Com isso, temos os resultados mais relevantes deste experimento para a proposta

Figura 15 – Matriz de confusão média para 10 janelas de 1 segundo com NB e HT.

		Classe Predita	
		Com Rainha	Sem Rainha
Classe Real	Com Rainha	VP 1885	FN 65 Erro do Tipo II
	Sem Rainha	FP 8 Erro do Tipo I	VN 382

Naive Bayes

		Classe Predita	
		Com Rainha	Sem Rainha
Classe Real	Com Rainha	VP 1894	FN 56 Erro do Tipo II
	Sem Rainha	FP 18 Erro do Tipo I	VN 372

Hoeffding Tree

Fonte: elaborado pelo autor (2023).

deste estudo e dentre os classificadores incrementais propostos, o NB é o mais adequado para a tarefa. Produzindo 0,93 s de tempo de execução com 97 % de taxa de acerto e com baixíssima incidência de erros do tipo I.

4.2 Resultados do experimento com colônias de abelhas sem ferrão (*Tetragonisca fiebrigi*)

A estrutura dos resultados no experimento com as abelhas jataí (sem ferrão) se deu de forma um pouco diferente. Primeiramente, como já falado na seção anterior, foi feita uma análise dos agrupamentos quanto a sua similaridade e dissimilaridade entre as classes e cada agrupamento de interesse. Dentre os números de agrupamentos possíveis, os de maior interesse na perspectiva dos dados são os números 2 (dois) e 6 (seis). Naturalmente, estes números possuem um maior relevância por representar formas diferentes de separar ou classificar os dados obtidos. Isto porque temos 6 colmeias em estudo em um delineamento binário onde três colmeias estavam com rainha no momento da gravação e as outras três não estavam. Então é necessário analisar estes dois agrupamentos prioritariamente para entender quais resultados das métricas de similaridade/dissimilaridade prevalecem. Desta maneira, é possível determinar se as individualidades entre cada uma das seis colmeias são mais expressivas na gravação do áudio ou se a presença de rainha teria mais expressão no padrão bioacústico encontrado.

Para tal, foram analisadas três métricas para os agrupamentos. O coeficiente *silhouette*, o índice de Calinski-Harabasz e o índice de Davies-Bouldin. Dividindo nas duas possíveis

Tabela 4 – Resultado da análise com dois ou seis agrupamentos.

Número de agrupamentos	<i>Silhouette</i>	Calinski-Harabasz	Davies-Bouldin
2	0,026	7,813	6,632
6	-0,013	4,489	8,302

Fonte: elaborado pelo autor (2023).

formas de agrupar os dados, com seis classes (uma para cada colônia) ou com duas classes (colônias com e sem rainha) obtivemos os resultados mostrados na tabela 4. Estes resultados mostram que para os três índices analisados, a utilização de dois agrupamentos é mais consistente. Nos levando a acreditar que as diferenças e similaridades intra e entre classes, relativas a presença de rainha, sobrepõem quaisquer outros padrões individuais de cada uma das colmeias.

Tabela 5 – Resultado da análise com o k-médias com $2 \leq k \leq 6$.

k (número de <i>clusters</i>)	<i>Silhouette</i>	Calinski-Harabasz	Davies-Bouldin
2	0,284	71000	1,325
3	0,250	63500	1,466
4	0,217	52800	1,551
5	0,222	46500	1,414
6	0,208	42800	1,413

Fonte: elaborado pelo autor (2023).

Utilizando o algoritmo k-médias para *clusterização*, variando o valor de k entre dois e seis *clusters*, obtivemos os resultados mostrados na tabela 5. Novamente, os dados das métricas mostram que o melhor número de agrupamentos para o problema são exatamente 2. Estes resultados reforçam ainda mais a hipótese de que a presença de rainha influencia com predominância nos padrões bioacústicos em relação às possíveis assinaturas individuais de cada colônia (e.g. número de abelhas, atividade forrageira, quantidade de cria, alimento). Isto significa que a assinatura bioacústica da presença de rainha tem forte relevância mesmo comparando entre colmeias e colônias diferentes. Estes são resultados preliminares que servem como base para a adoção de métodos supervisionados para o problema em questão.

As próximas seções tratam exatamente do problema supervisionado. Relacionando os resultados dos três classificadores escolhidos: o *Multilayer Perceptron* (Seção 4.2.1), a *Extreme Learning Machine* (Seção 4.2.2) e o *AdaBoost* (Seção 4.2.3), respectivamente. Ao final, uma comparação geral entre os três classificadores (Seção 4.2.4).

4.2.1 Multilayer Perceptron

Os resultados da utilização da rede neural MLP para o experimento em questão são relacionados na Tabela 6. A tabela dispõe o resultado de todas as métricas analisadas em cada uma das 10 iterações. O tempo e acurácia são destacados porque são os fatores que foram utilizados para comparação entre os modelos.

Tabela 6 – Resultado das métricas do MLP.

Iteração	Multilayer Perceptron				
	Tempo (s)	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
1	170,8	93,5	92,8	94,0	93,0
2	180,0	93,3	93,5	92,7	93,3
3	158,4	93,5	93,1	93,6	93,2
4	168,3	93,3	94,2	92,0	93,7
5	189,9	93,5	93,1	93,6	93,2
6	175,7	93,4	93,4	93,2	93,3
7	182,5	93,5	92,8	93,9	93,0
8	163,7	93,1	92,3	93,8	92,6
9	162,2	93,1	93,0	92,8	93,0
10	186,1	93,3	93,0	93,2	93,1
média	173,3	93,4	93,1	93,4	93,1

Fonte: elaborado pelo autor (2023).

É possível notar que, com o MLP, temos um tempo de treino/teste médio de 173,3 segundos. Como estamos lidando com modelos em *batch* ou *offline* neste experimento de abelhas jataí, vale ressaltar que os tempos são diferentes do experimento anterior e devem ser analisados sobre a perspectiva da abordagem equivalente. A acurácia média do MLP ficou acima dos 93%, bem como as médias das demais métricas. Desta forma o MLP demonstra ter uma estabilidade na acurácia e nas demais métricas.

4.2.2 Extreme Learning Machine

Os resultados com a rede neural ELM para o experimento em questão são relacionados na Tabela 7. A tabela dispõe o resultado de todas as métricas analisadas em cada uma das 10 iterações. O tempo e acurácia são destacados porque são os fatores que foram utilizados para comparação entre os modelos.

Utilizando o ELM, temos um tempo de treino/teste médio de 66,8 segundos. A acurácia do ELM ficou acima dos 89%. As demais métricas variaram um pouco, ficando acima de 88% mas o recall médio chegando na faixa dos 90%, indicando uma menor incidência de erros do tipo II. Em comparação com o MLP, houve uma redução drástica no tempo de treino/teste, em

Tabela 7 – Resultado das métricas do ELM.

Extreme Learning Machine					
Iteração	Tempo (s)	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
1	67,8	89,4	88,7	90,9	89,1
2	67,0	89,3	88,5	90,9	89,0
3	66,8	89,2	88,5	90,7	89,0
4	66,7	89,3	88,4	91,0	88,9
5	65,2	89,4	88,7	90,9	89,1
6	72,8	89,4	88,6	90,9	89,1
7	66,7	89,6	88,8	91,1	89,2
8	72,0	88,9	88,3	90,4	88,7
9	66,3	89,1	88,6	90,4	88,9
10	66,7	89,3	88,5	90,9	89,0
média	66,8	89,3	88,5	90,9	89,0

Fonte: elaborado pelo autor (2023).

contrapartida, a acurácia também diminuiu. Desta forma, se o desempenho for mais importante do que alguns pontos percentuais de acurácia então o ELM seria mais adequado para a situação, do contrário, o MLP seria o mais indicado.

4.2.3 AdaBoost

Os resultados com o ADB para o experimento em questão são relacionados na Tabela 8. A tabela dispõe o resultado de todas as métricas analisadas em cada uma das 10 iterações. O tempo e acurácia são destacados porque são os fatores que foram utilizados para comparação entre os modelos.

Tabela 8 – Resultado das métricas do ADB.

AdaBoost					
Iteração	Tempo (s)	Acurácia (%)	Precisão (%)	Recall (%)	$F_{\beta=0,5}$ (%)
1	351,2	87,4	87,2	86,9	87,1
2	351,4	87,1	87,0	86,7	86,9
3	350,5	87,1	86,9	86,6	86,9
4	351,2	87,3	87,1	86,8	87,1
5	350,1	87,7	87,5	87,3	87,5
6	359,9	87,4	87,5	86,7	87,4
7	359,2	87,6	87,5	87,1	87,4
8	350,2	86,9	86,6	86,8	86,6
9	357,1	87,0	87,0	86,3	86,9
10	357,0	87,2	87,1	86,8	87,0
média	351,3	87,2	87,1	86,8	87,1

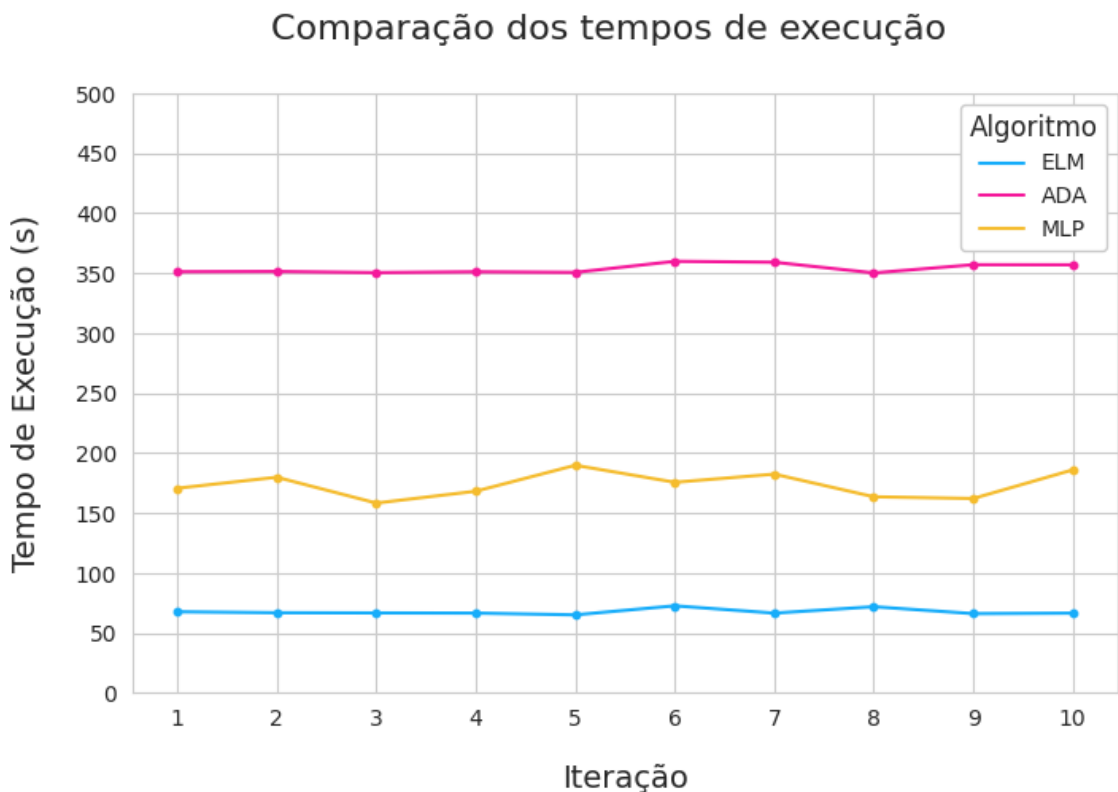
Fonte: elaborado pelo autor (2023).

Vemos que, com o ADB, temos um tempo de treino/teste médio de 321,3 segundos. A acurácia do ADB ficou um pouco acima dos 87%. As demais métricas variaram um pouco, ficando acima de 87% mas o recall médio parando na faixa dos 86%, indicando uma maior incidência de erros do tipo II. Em comparação com os demais classificadores, o ADB tem um

desempenho inferior aos demais. Isto é, o tempo de treino/teste é bem superior aos demais e a acurácia é também bem inferior aos demais. Sendo assim, o ADB acaba sendo um classificador que não está em pé de igualdade e nem dentro de uma possível faixa de comparação ou indicação em casos mais específicos dentro da perspectiva deste experimento.

4.2.4 Comparando os 3 classificadores

Figura 16 – Tempo de execução entre os classificadores.



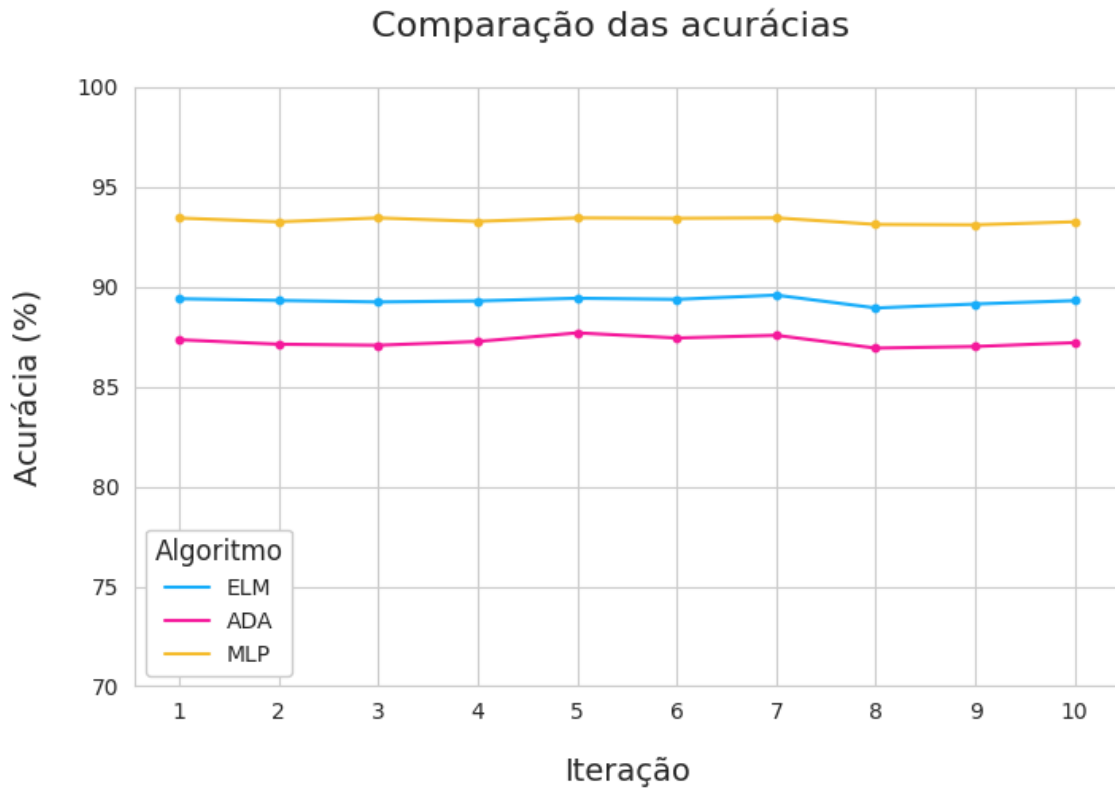
Fonte: elaborado pelo autor (2023).

A Figura 16 relaciona os tempos de execução de cada um dos três classificadores *in batch* para cada uma das 10 iterações. É possível notar no gráfico que o ADB apresentou um tempo de execução maior e mais distante dos demais. Enquanto isso, a ELM tem um melhor desempenho seguido do MLP que se encontra na faixa intermediária, contudo, mais próximo à curva do ELM.

No que diz respeito a acurácia dos classificadores, temos a Figura 10 com a relação entre a acurácia observada em cada iteração. Para a acurácia, observa-se uma disposição diferente. Nesta métrica o MLP obteve os melhores resultados. Seguido pelo ELM e o ADB respectivamente. Entretanto, o ELM e o ADB encontram-se mais próximos um do outro do que

do MLP. Neste aspecto, ele tem um destaque positivo na acurácia.

Figura 17 – Acurácia entre os classificadores.



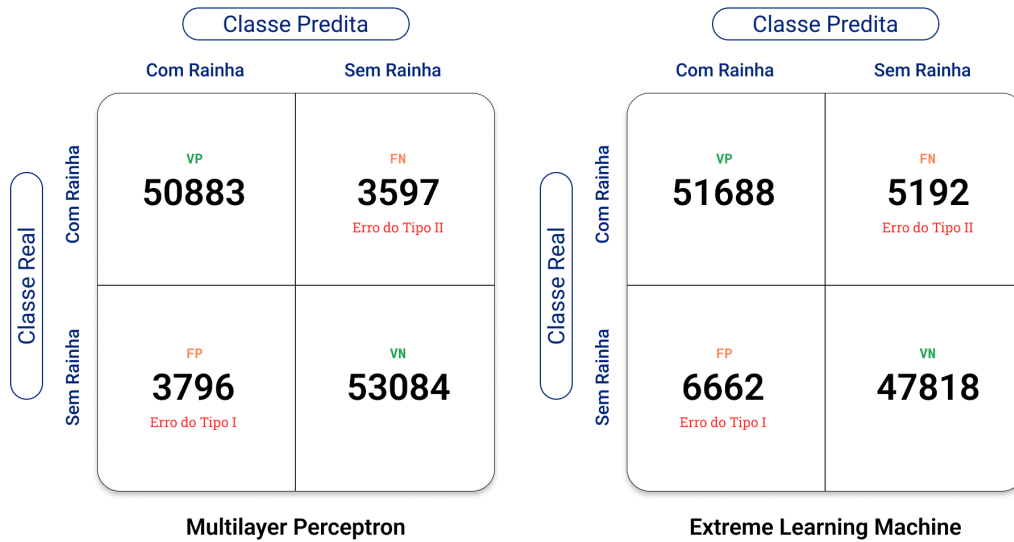
Fonte: elaborado pelo autor (2023).

Analisando de forma geral, enquanto o ELM tem o destaque positivo no tempo de execução, o MLP se destaca na acurácia. E os resultados da AART mostraram que os tempos e as acurácias são diferentes entre os classificadores. Ou seja, cada classificador produz tempo e acurácia distintas dos demais sob análise. Entretanto, o ADB não foi superior em nenhum dos casos: teve o tempo de execução mais custoso e a acurácia mais baixa. Por esta razão, não será mais considerado na comparação uma vez que não é mais qualificado que os demais para nenhuma das métricas sob estudo.

Na perspectiva das duas métricas em análise de desempenho (tempo de execução e acurácia) temos a ELM superior em uma e o MLP em outra. Expandindo para as outras métricas do classificador, o MLP possui precisão, *recall* e $F_{\beta=0,5}$ superiores ao ELM. Isto indica uma menor incidência de erros e principalmente erros do tipo I. Essa constatação é reforçada pela matriz de confusão média entre os dois classificadores *in batch*, mostrado na Figura 18. Sob o ponto de vista deste estudo, o MLP possui um desempenho melhor que supera o déficit em tempo de execução. Isto é, para o problema em estudo, é considerável esperar um pouco mais em tempo de execução para a aquisição de métricas mais elevadas e adequadas. Até porque o

aumento no tempo de execução não é tão elevado embora seja considerável. Ainda assim, para alguns problemas o ELM pode ser o mais adequado devido seu baixo tempo de execução.

Figura 18 – Matriz de confusão média para 200 amostras com MLP e ELM.



Fonte: elaborado pelo autor (2023).

Desta forma, o classificador *in batch* ELM se mostrou ser o mais indicado quando o tempo for o fator mais limitante do que as métricas de avaliação do classificador como acurácia, precisão, etc. Mas no caso das métricas serem os fatores mais sensíveis e o tempo de treino/teste do classificador puder ser um pouco mais flexível, que é o caso deste estudo, o MLP possui o melhor desempenho. Para a análise específica deste estudo, recomendamos a utilização do classificador *in batch* MLP. Esperando alguns segundos a mais é possível ter um retorno de métricas mais eficientes além de uma bem menor incidência de erros, principalmente erros do tipo I.

4.3 Discussão dos resultados

Na análise de desempenho realizada neste trabalho, levamos em consideração duas métricas de avaliação: tempo de execução e acurácia. O tempo de execução é importante por ser mais desejável uma identificação rápida de cenários como ausência de rainha para que a tomada de decisão por parte do apicultor/meliponicultor seja igualmente rápida. A rapidez na identificação e na posterior tomada de decisão pode ser crucial para o manejo das abelhas e na garantia do seu bem-estar. Já a acurácia é o quão eficiente o classificador é, ou em outras palavras,

o quanto o classificador acerta suas predições para cada amostra. É desejável que a acurácia seja a maior possível uma vez que isso significa que o classificador está identificando um maior número de amostras corretamente e conseqüentemente está identificando de forma confiável e sem erros os cenários de presença/ausência de rainha para o manejador da(s) colônia(s).

Entretanto, estas duas métricas precisam de uma análise de *trade-off* ou ponderamento entre elas para uma melhor decisão para o determinado problema. Isto é, não se deseja que o modelo de classificação seja muito rápido em troca de muitos erros de classificação e nem que ele acerte todas as classificações mas em troca de um tempo de execução tão demorado que a tomada de decisão do manejador seja atrasada a ponto da colônia entrar em colapso ou em grande risco. Desta forma, existe um *trade-off*, uma troca ponderada, entre as métricas de forma que as duas possuam valores aceitáveis e aplicáveis para o problema de monitoramento de colônias de abelhas. Para isso foi utilizado o teste estatístico descrito na Seção 3.4. Se apenas acurácia tivesse impacto para a eficiência da mitigação do problema, o RF seria o mais indicado por ter acurácias mais elevadas que os demais classificadores (Figura 14). Contudo, o tempo de execução é muito maior que os demais em troca de um aumento significativo estatisticamente, mas não tão grande assim para o problema em questão. Fazendo o *trade-off* adequado, vemos que apenas o HT e o NB podem ser escolhidos como classificadores adequados para o problema por possuírem as duas métricas: tempo (Figura 13) e acurácia (Figura 14) dentro de um padrão aceitável ao mesmo tempo. Dentre os dois (HT e NB), como as métricas são bem parecidas, com o NB sendo levemente mais eficiente no tempo de execução e HT levemente mais eficiente na acurácia, as outras métricas foram levadas em consideração. Desta maneira, como o NB tem uma precisão e medida $F_{\beta=0,5}$ superior ao HT (Figura 15) e conseqüentemente uma menor incidência de erros do tipo I, o NB fica recomendado neste trabalho como o melhor classificador incremental dentre os adotados para o problema abordado no primeiro experimento deste trabalho.

A hipótese nula do teste AART é que as distribuições analisadas advém de uma mesma distribuição original e, desta forma, são consideradas distribuições iguais. Sendo assim, o cenário ideal seria uma combinação de fatores onde os tempos de execução são menores que os demais mas acurácias são iguais estatisticamente para as demais combinações de fatores. Ou, em outras palavras, para um determinado classificador, uma combinação de fatores com menor tempo de execução não necessariamente registre acurácias estatisticamente diferentes e inferiores quando comparado com combinações com tempos de execução maiores. Dentre os classificadores incrementais, apenas o NB esteve no cenário ideal. Utilizando 10 janelas

de 1 segundo, o NB registrou um tempo médio de 0,93 s (Figura 11) que é menor e diferente estatisticamente das demais combinações. Não obstante, registrou uma acurácia de 97 % (Tabela 2) que, de acordo com a AART, não rejeitou-se a hipótese nula e consequentemente possui distribuições semelhantes às acurácias das demais combinações de fatores. Ou seja, escolher 10 janelas de 1 segundo como combinação de fatores para o NB tem um impacto positivo no tempo de execução e não gera nenhum impacto na acurácia. Sendo, portanto, o caso ideal para a taxa de amostragem de um classificador nesta análise de desempenho.

Já no caso das abelhas jataí, temos uma situação parecida onde um dos classificadores não possui resultados satisfatórios para uma das métricas, no caso o ADB, possui um tempo de execução superior aos demais (Figura 16) e ainda assim uma acurácia inferior (Figura 17). Portanto, obteve-se resultados insuficientes em comparação aos demais para ambas as métricas de avaliação consideradas neste estudo. Em contrapartida os outros dois classificadores *in batch* utilizados (MLP e ELM) obtiveram resultados superiores para cada uma das respectivas métricas. Neste caso, o MLP é o melhor classificador em acurácia, enquanto o ELM é superior na métrica tempo de execução. Em uma perspectiva de outros problemas, a decisão entre os dois poderia se dar dependendo da métrica mais relevante para o problema lidado. Se para o problema em questão o tempo de execução for a métrica mais sensível, recomenda-se o ELM. Do contrário, se a acurácia for a métrica mais sensível, o MLP é o recomendado. No caso deste trabalho, o *trade-off* entre as métricas tende à acurácia. Em troca de um pouco mais de custo em tempo de execução, se obtém diferenças estatisticamente significantes e superiores na acurácia utilizando o MLP. Além do mais, foram analisadas as outras métricas e como o MLP possui uma maior precisão e medida $F_{\beta=0,5}$ e consequentemente uma menor incidência de erros do tipo I (Figura 18), se reafirma como o modelo classificador recomendado para o segundo experimento abordado neste trabalho.

4.4 Sumário do Capítulo

Neste capítulo foram tratados os resultados obtidos através da metodologia experimental descrita no Capítulo 3. Os resultados foram relacionados sumariamente em torno dos dois experimentos principais com abelhas com ferrão e sem ferrão. Na Seção 3.2.1 temos os resultados do primeiro experimento com abelhas melíferas africanizadas e classificadores incrementais, passando por cada um dos classificadores incrementais utilizados: HT (Seção 4.1.1), NB (Seção 4.1.2) e RF (Seção 4.1.3) e em seguida comparando os resultados entre eles

(Seção 4.1.4). De maneira análoga, foram abordados os resultados do segundo experimento com abelhas jataí na Seção 4.2. Da mesma forma, passando por cada classificador *in batch* utilizado na experimentação: MLP (Seção 4.2.1), ELM (Seção 4.2.2) e ADB (Seção 4.2.3) e então comparando os resultados entre eles (Seção 4.2.4). Ao final do capítulo, na Seção 4.3, realizamos uma breve discussão a respeito dos resultados obtidos.

5 CONCLUSÃO

Neste capítulo, apresentamos as conclusões desta dissertação e sugerimos algumas perspectivas para a sequência do trabalho. Na Seção 5.1, relacionamos as conclusões às questões de pesquisa e hipóteses (Seção 1.3) as quais, através de evidências suportadas pelos resultados (Capítulo 4), atingem os objetivos (Seção 1.4). Na sequência, a Seção 5.2 traz as publicações científicas realizadas durante o período deste mestrado. Por fim, na Seção 5.3 sugerimos algumas possibilidades de trabalhos futuros.

5.1 Contribuições e Objetivos Alcançados

A principal contribuição desta dissertação, diretamente associada ao seu **Objetivo geral**, é a possibilidade de identificar de forma rápida e acurada (vide Figuras 13, 14, 16 e 17) e também de forma precisa (vide Tabelas 1, 2, 3, 6, 7 e 8) a presença de rainha em colônias de abelhas de espécies comuns no Brasil utilizando seus padrões bioacústicos. Utilizamos duas espécies de abelhas: melíferas africanizadas (*Apis mellifera*), cuja presença é típica na região Nordeste e a jataí (*Tetragonisca fiebrigi*), cuja presença é típica na região Sul do Brasil. Desta maneira, conforme as evidências destacadas na Tabela 9, confirmamos a **Hipótese principal** desta dissertação.

Uma contribuição secundária é a identificação de maneira rápida (0,93 s), acurada (97%) e precisa (99,7%) da presença de rainha em colmeias de abelhas melíferas (*Apis mellifera*) do tipo africanizada utilizando classificadores incrementais (vide Tabelas 1, 2, 3). Assim, confirmamos a **Hipótese #1** associada à **QP #1** e, conseqüentemente, ao **Objetivo específico 1** da Tabela 9. O melhor resultado foi alcançado pelo classificador NB (Seção 4.1.4). Estes resultados foram obtidos utilizando uma taxa de amostragem bioacústica de apenas 10 janelas de 1 segundo por dia (Tabela 2). Além disso, os fatores (quantidade e tamanho das janelas) são estatisticamente significativos para as métricas dos classificadores. Ainda assim, com 10 janelas de 1 segundo, é possível ter um tempo de execução bem menor em troca de poucas casas decimais de acurácia ou até mesmo sem impacto na acurácia (NB), conforme evidenciado na Seção 4.1.2. A partir desta taxa, o apicultor pode obter uma resposta rápida e precisa da presença de rainha nas colônias de abelhas melíferas sob seu manejo. Esta rápida identificação aliada ao baixo custo de recursos computacionais assegura uma tomada de decisão ágil para garantir o bem-estar da colônia monitorada. Sendo assim, a **Hipótese #2** relacionada à **QP #2** e, portanto,

ao **Objetivo específico 2** (Tabela 9), é também validada.

Uma terceira contribuição desta dissertação diz respeito às colônias de abelhas jataí (*Tetragonisca fiebrigi*). Mesmo com as peculiaridades de cada colônia, os agrupamentos relativos às classes com e sem rainha mostraram-se mais adequados (vide Tabelas 4 e 5). Além disso, foi possível identificar a presença de rainha nas colônias através do padrão bioacústico de forma acurada (93%) e precisa (92%) com o classificador MLP. Há padrões bioacústicos únicos e predominantes relativos à presença da rainha, o que nos permite validar a **Hipótese #3** relacionada à **QP #3** e ao **Objetivo específico 3** (Tabela 9).

Finalmente, os algoritmos de classificação convencionais mostraram-se eficientes ($\geq 87\%$ de acurácia, considerando todos os classificadores *in batch* analisados) mas não tanto quanto no caso dos classificadores incrementais ($\geq 96\%$ de acurácia, considerando todos os classificadores incrementais analisados) para o monitoramento de fluxos de dados em tempo real de uma única colônia. Concluímos, portanto, que detectar a presença da rainha por colmeia usando monitorando em tempo real é melhor que fazê-lo entre colmeias. As métricas de avaliação de cada abordagem são evidenciadas nas Figuras 13 e 14 para a classificação incremental das abelhas com ferrão e nas Figuras 16 e 17 para a classificação *in batch* das abelhas sem ferrão. Assim sendo, podemos validar a **Hipótese #4** relacionada à **QP #4** referente ao **Objetivo específico 4** na Tabela 9.

Tabela 9 – Objetivos (geral e específicos) e evidências dos seus alcances.

Objetivos	Evidências do alcance
Objetivo geral	Figuras 13, 14, 16 e 17 e nas Tabelas 1, 2, 3, 6, 7 e 8
Objetivo específico 1	Tabelas 1, 2, 3
Objetivo específico 2	Tabelas 1, 2, 3
Objetivo específico 3	Tabelas 4 e 5 e Figuras 16 e 17
Objetivo específico 4	Figuras 13, 14 e 16, 17

Fonte: elaborado pelo autor (2023).

5.2 Publicações

Publicamos dois artigos científicos completos no Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA) ¹, evento satélite do Congresso da Sociedade Brasileira de Computação:

¹ <https://sol.sbc.org.br/index.php/wcama>

- **RODRIGUES, Ícaro de Lima;** MELO, Davyd B. de; SILVA, Daniel de Amaral da; RYBARCZYK, Yves; GOMES, Danielo G.. Padrões Bioacústicos como Identificadores Precisos da Presença de Rainha em Colmeias de Abelhas Melíferas. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 13. , 2022, Niterói. Porto Alegre: Sociedade Brasileira de Computação, 2022 . p. 11-20. ISSN 2595-6124. DOI: <https://doi.org/10.5753/wcama.2022.222913> (**prêmio de melhor artigo**);
- **RODRIGUES, Ícaro de Lima;** MELO, Davyd B. de; FREITAS, Breno M.; GOMES, Danielo G.. Detecção de Anomalias em Padrões Acústicos, de Temperatura e Umidade Sazonais para Abelhas Melíferas (*Apis mellifera* L.). In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA), 12. , 2021, Evento Online. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 69-78. ISSN 2595-6124. DOI: <https://doi.org/10.5753/wcama.2021.15738>.

5.3 Algumas Perspectivas

Muito embora o objeto de estudo desta dissertação tenha sido somente as abelhas melíferas africanizadas (*Apis mellifera*) e abelhas jataí (*Tetragonisca fiebrigi*), concluímos em perspectiva para outras espécies de abelhas. Acreditamos que os métodos propostos nesta dissertação podem ser reproduzidos e aplicados para quaisquer espécies de abelhas, com ou sem ferrão, inclusive as comuns do Brasil, sendo possível identificar e alertar apicultores /meliponicultores sobre outros tipos de anomalias fora do padrão bioacústico natural da colônia (e.g. intoxicação, infestação de parasitas, ataque de predadores).

Outro possível trabalho futuro seria aplicar nossas contribuições ao processo de provisionamento e postura nas células de cria (POP, do inglês *cell provisioning and oviposition process*), fenômeno característico das abelhas sem ferrão. Dado que a frequência do POP prediz aumento ou queda populacional, seria interessante correlacionar vídeos internos com áudios ocorrendo nos momentos do POP para diagnosticar e prever a saúde dos enxames.

Um leque mais amplo de perspectiva seria estender este trabalho para outras espécies da fauna (e.g. morcegos, aves) sobretudo considerando listas vermelhas de espécies ameaçadas de extinção².

² <https://www.sema.ce.gov.br/lista-vermelha-de-especies-ameacadas-da-fauna-do-ceara/>

REFERÊNCIAS

- ABDOLLAHI, M.; GIOVENAZZO, P.; FALK, T. H. Automated beehive acoustics monitoring: A comprehensive review of the literature and recommendations for future work. **Applied Sciences**, MDPI, v. 12, n. 8, p. 3920, 2022.
- ARIAS-AGUILAR, A.; HINTZE, F.; AGUIAR, L. M.; RUFRAY, V.; BERNARD, E.; PEREIRA, M. J. R. Who's calling? acoustic identification of brazilian bats. **Mammal Research**, Springer, v. 63, p. 231–253, 2018.
- BARTH, F. G.; HRNCIR, M.; JARAU, S. Signals and cues in the recruitment behavior of stingless bees (meliponini). **Journal of Comparative Physiology A**, Springer, v. 194, n. 4, p. 313–327, 2008.
- BERRAR, D. Bayes' theorem and naive bayes classifier. **Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics**, Elsevier Science Publisher Amsterdam, The Netherlands, v. 403, p. 412, 2018.
- BRAGA, A. R. Modelos de classificação para predição do bem estar de colônias da abelha apis mellifera. 2020.
- BRAGA, A. R.; FREITAS, B. M.; GOMES, D. G.; BEZERRA, A. D.; CAZIER, J. A. Forecasting sudden drops of temperature in pre-overwintering honeybee colonies. **Biosystems Engineering**, v. 209, p. 315–321, 2021. ISSN 1537-5110. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1537511021001665>>.
- 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.
- BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.
- BROWN, M. J.; DICKS, L. V.; PAXTON, R. J.; BALDOCK, K. C.; BARRON, A. B.; CHAUZAT, M.-P.; FREITAS, B. M.; GOULSON, D.; JEPSEN, S.; KREMEN, C. *et al.* A horizon scan of future threats and opportunities for pollinators and pollination. **PeerJ**, PeerJ Inc., v. 4, p. e2249, 2016.
- CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974.
- CHENGSHENG, T.; HUACHENG, L.; BING, X. Adaboost typical algorithm and its application research. In: EDP SCIENCES. **MATEC Web of Conferences**. [S.l.], 2017. v. 139, p. 00222.
- DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, n. 2, p. 224–227, 1979.
- DINEVA, K.; ATANASOVA, T. *et al.* Osemn process for working over data acquired by iot devices mounted in beehives. **Curr. Trends Nat. Sci**, v. 7, n. 13, p. 47–53, 2018.
- DING, S.; XU, X.; NIE, R. Extreme learning machine and its applications. **Neural Computing and Applications**, Springer, v. 25, n. 3, p. 549–556, 2014.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of computer and system sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GIANNAKOPOULOS, T. pyaudioanalysis: An open-source python library for audio signal analysis. **PLOS ONE**, Public Library of Science, v. 10, n. 12, p. 1–17, 12 2015. Disponível em: <<https://doi.org/10.1371/journal.pone.0144610>>.

GOMES, H. M.; BIFET, A.; READ, J.; BARDDAL, J. P.; ENEMBRECK, F.; PFHARINGER, B.; HOLMES, G.; ABDESSALEM, T. Adaptive random forests for evolving data stream classification. **Machine Learning**, Springer, v. 106, n. 9, p. 1469–1495, 2017.

HADJUR, H.; AMMAR, D.; LEFÈVRE, L. Toward an intelligent and efficient beehive: A survey of precision beekeeping systems and services. **Computers and Electronics in Agriculture**, v. 192, p. 106604, 2022. ISSN 0168-1699. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0168169921006219>>.

HASTIE, T.; ROSSET, S.; ZHU, J.; ZOU, H. Multi-class adaboost. **Statistics and its Interface**, International Press of Boston, v. 2, n. 3, p. 349–360, 2009.

HOWARD, D.; DURAN, O.; HUNTER, G.; STEBEL, K. Signal processing the acoustics of honeybees (*apis mellifera*) to identify the "queenless" state in hives. **Proceedings of the Institute of Acoustics**, v. 35, p. 290–297, 01 2013.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE. **2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)**. [S.l.], 2004. v. 2, p. 985–990.

HULTEN, G.; SPENCER, L.; DOMINGOS, P. Mining time-changing data streams. In: **Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2001. p. 97–106.

HUNT, J.; RICHARD, F.-J. Intracolony vibroacoustic communication in social insects. **Insectes Sociaux**, Springer, v. 60, n. 4, p. 403–417, 2013.

HUNTER, D. H. O. D. G.; STEBEL, K. Signal processing the acoustics of honeybees (*apis mellifera*) to identify the 'queenless' state in hives. **Proceedings of the Institute of Acoustics**, v. 35, n. 1, p. 290, 2013.

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 B: biological sciences**, The Royal Society London, v. 274, n. 1608, p. 303–313, 2007.

LAUTENBACH, S.; SEPPELT, R.; LIEBSCHER, J.; DORMANN, C. F. Spatial and temporal trends of global pollination benefit. **PLoS one**, Public Library of Science, v. 7, n. 4, p. e35954, 2012.

- LENG, Q.; QI, H.; MIAO, J.; ZHU, W.; SU, G. One-class classification with extreme learning machine. **Mathematical problems in engineering**, Hindawi, v. 2015, 2015.
- 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**, Springer, v. 48, n. 5, p. 666–680, 2017.
- NA, S.; XUMIN, L.; YONG, G. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In: IEEE. **2010 Third International Symposium on intelligent information technology and security informatics**. [S.l.], 2010. p. 63–67.
- NGUYEN, H.-L.; WOON, Y.-K.; NG, W.-K. A survey on data stream clustering and classification. **Knowledge and information systems**, Springer, v. 45, n. 3, p. 535–569, 2015.
- NIEH, J. C.; ROUBIK, D. W. Potential mechanisms for the communication of height and distance by a stingless bee, *melipona panamica*. **Behavioral Ecology and Sociobiology**, Springer, v. 43, p. 387–399, 1998.
- NOLASCO, I.; TERENCEZI, A.; CECCHI, S.; ORCIONI, S.; BEAR, H. L.; BENETOS, E. Audio-based identification of beehive states. In: **ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2019. p. 8256–8260. ISSN 2379-190X.
- OLLERTON, J.; WINFREE, R.; TARRANT, S. How many flowering plants are pollinated by animals? **Oikos**, Wiley Online Library, v. 120, n. 3, p. 321–326, 2011.
- 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. *et al.* Safeguarding pollinators and their values to human well-being. **Nature**, Nature Publishing Group, v. 540, n. 7632, p. 220–229, 2016.
- RAMCHOUN, H.; GHANOU, Y.; ETTAOUIL, M.; IDRISSE, M. A. J. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence . . .*, 2016.
- RISH, I. *et al.* An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.
- ROBLES-GUERRERO, A.; SAUCEDO-ANAYA, T.; GONZÁLEZ-RAMÉREZ, E.; GALVÁN-TEJADA, C. E. Frequency analysis of honey bee buzz for automatic recognition of health status: A preliminary study. **Res. Comput. Sci.**, v. 142, p. 89–98, 2017.
- ROBLES-GUERRERO, A.; SAUCEDO-ANAYA, T.; GONZÁLEZ-RAMÍREZ, E.; ROSA-VARGAS, J. I. D. L. 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**, Elsevier, v. 159, p. 69–74, 2019.
- RODRIGUES Ícaro; MELO, D.; SILVA, D.; RYBARCZYK, Y.; GOMES, D. Padrões bioacústicos como identificadores precisos da presença de rainha em colmeias de abelhas melíferas. In: **Anais do XIII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**. Porto Alegre, RS, Brasil: SBC, 2022. p. 11–20. ISSN 2595-6124. Disponível em: <<https://sol.sbc.org.br/index.php/wcama/article/view/20692>>.

- ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.
- ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.
- SILVA, D.; RODRIGUES Ícaro; BRAGA, A.; NOBRE, J.; FREITAS, B.; GOMES, D. An autonomic, adaptive and high-precision statistical model to determine bee colonies well-being scenarios. In: **Anais do XI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**. Porto Alegre, RS, Brasil: SBC, 2020. p. 31–40. ISSN 2595-6124. Disponível em: <<https://sol.sbc.org.br/index.php/wcama/article/view/11017>>.
- SINAGA, K. P.; YANG, M.-S. Unsupervised k-means clustering algorithm. **IEEE access**, IEEE, v. 8, p. 80716–80727, 2020.
- STOWELL, D.; PLUMBLEY, M. D. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. **PeerJ**, PeerJ Inc., v. 2, p. e488, 2014.
- SÁNCHEZ-BAYO, F.; WYCKHUYS, K. Worldwide decline of the entomofauna: A review of its drivers. **Biological Conservation**, v. 232, 01 2019.
- WILK, J. T.; BĄK, B.; ARTIEMJEW, P.; WILDE, J.; SIUDA, M. Classifying the biological status of honeybee workers using gas sensors. **Sensors**, v. 21, n. 1, 2021. ISSN 1424-8220. Disponível em: <<https://www.mdpi.com/1424-8220/21/1/166>>.
- WINSTON, M. L. **The biology of the honey bee**. [S.l.]: harvard university press, 1991.
- WOBBROCK, J. O.; FINDLATER, L.; GERGLE, D.; HIGGINS, J. J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: **Proceedings of the SIGCHI conference on human factors in computing systems**. [S.l.: s.n.], 2011. p. 143–146.
- ZACEPINS, A.; BRUSBARDIS, V.; MEITALOVŠ, J.; STALIDZANS, E. Challenges in the development of precision beekeeping. **Biosystems Engineering**, v. 130, p. 60–71, 2015. ISSN 1537-5110. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1537511014002086>>.
- ZGANK, A. Bee swarm activity acoustic classification for an iot-based farm service. **Sensors**, v. 20, n. 1, 2020. ISSN 1424-8220. Disponível em: <<https://www.mdpi.com/1424-8220/20/1/21>>.
- Zhu, W.; Miao, J.; Qing, L. Constrained extreme learning machine: A novel highly discriminative random feedforward neural network. In: **2014 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2014. p. 800–807.
- ŽGANK, A. Acoustic monitoring and classification of bee swarm activity using mfcc feature extraction and hmm acoustic modeling. In: **2018 ELEKTRO**. [S.l.: s.n.], 2018. p. 1–4.