IoT para o Desenvolvimento: Construindo um Algoritmo de Classificação para Ajudar os Apicultores a Detectar Prematuramente Problemas de Saúde de Abelhas

IoT for Development: Building a Classification Algorithm to Help Beekeepers Detect Honeybee Health Problems Early

Completed Research

Antonio Rafael Braga
Universidade Federal do Ceará
rafaelbraga@ufc.br

Edgar E. Hassler
Appalachian State University
hassleree@appstate.edu

Danielo G. Gomes
Universidade Federal do Ceará
danielo@ufc.br

Breno M. Freitas
Universidade Federal do Ceará
freitas@ufc.br

Joseph A. Cazier
Appalachian State University
cazierja@appstate.edu

Abstract

Bees are the main pollinators of most wild plant species and are essential for the maintenance of plant ecosystems and for food production. However, in recent years they are suffering from deforestation and pesticides. Here, we propose a method to identify the health status of *Apis mellifera* colonies. We trained, validated and tested 4 classification algorithms (Naive Bayes, k-NN, Random Forest and Neural Networks) on actual data from a beehive that was monitored for 6 months. For the generation of the classification model, we take into account data from internal sensors to the hive (temperature, relative humidity, and weight), external data (temperature, pressure, wind speed, and rainfall). We also use data from inspections performed weekly by a specialist in beekeeping. We compared the four algorithms and arrived at a high precision classification model to automatically identify the health status of bee colonies.

Keywords

Precision beekeeping, *Apis mellifera*, data mining, classification.

Introdução

A *Washington State University Extension* estima que as abelhas fornecem anualmente apenas nos EUA um serviço de polinização no valor de $18 bilhões. Contudo, apesar do crescimento global do número de colônias domesticadas de *Apis mellifera*, a quantidade de abelhas vem diminuindo nos EUA desde a década de quarenta, e em alguns países da Europa desde a década de sessenta (Potts et al. 2009). De acordo com a *Yale School of Forestry and Environmental Studies*, os apicultores norte americanos perderam 30% de suas colônias a cada inverno de 2003 a 2013 (Jacobs et al 2017). No Brasil, nos últimos anos, os apicultores do semiárido do Nordeste enfrentaram severas perdas na produção de mel devido a

um processo biológico conhecido como abandono das colmeias (*absconding*). Nesse processo, todas as abelhas deixam o ninho estabelecido para buscar outro novo e mais adequado (Freitas et al. 2007).

No Nordeste do Brasil, onde altas temperaturas são comuns na maior parte do ano, um grande número de colônias são perdidas devido ao abandono das colmeias. Como resultado, a estiagem prolongada de 2012 levou 70% das colônias a abandonarem os ninhos, o que causou uma queda de 66% na produção de mel em relação a 2011 apenas no estado do Piauí (Kridi et al. 2014). No estado do Rio Grando Sul, no Brasil, nos últimos meses de 2018 foram registradas as perdas de pelo menos 1200 colônias, prejuízo que passa de R$ 1 milhão. Estudos indicam que as perdas podem estar associadas ao pesticida Fipronil, o produto é usado no Brasil para proteger sementes de soja (Florentino 2019).

Para tentar identificar com antecedência problemas nas colônias, geralmente o apicultor realiza um procedimento de inspeção que consiste na verificação visual da colônia por meio da abertura da colmeia. É através da inspeção visual que o apicultor pode detectar uma série de problemas, inclusive doenças. Contudo, esse tipo de verificação é estressante para a colônia, pois provoca um desequilíbrio do microclima dentro da colmeia, atrapalha o trabalho das abelhas forrageiras, das operárias responsáveis pelo trabalho interno do ninho, rompe a organização da colônia e pode matar operárias ou até mesmo a rainha esmagadas com a retirada e inserção dos quadros. Além de ser um processo invasivo, uma inspecção minuciosa demanda tempo, o que pode atrapalhar e até comprometer as funções de polinização e produção de mel das abelhas. Assim, qualquer intrusão nas atividades comuns das colônias deve ser reduzida a um mínimo e ocorrer apenas quando estritamente necessário. Logo, saber de antemão quando há necessidade de inspecionar uma colméia e quais são as ações a serem tomadas evitará a interferências desnecessárias nas colônias, as perdas de colônias e, consequentemente, às perdas financeiras.

Além disso, os apicultores parecem não ter técnicas padronizadas de manutenção de registros de inspeções, muitas vezes sem registros sistemáticos. Menos ainda realizam análises para determinar quais práticas são eficientes para resolver ou evitar problemas. Assim, estabelecendo um padrão para avaliação de saúde para as colônias de abelhas, podemos utilizar o enorme potencial em usar dados de inspeção de colônias para melhorar a apicultura em geral. Uma abordagem possível é a utilização de uma "Lista de verificação de saúde da colônia" do inglês "Healthy Colony Checklist" (HCC), que ajuda a agendar tarefas, atribuir recursos e acompanhar os resultados.

Por outro lado, a utilização e análise de sensores nas colmeias podem reduzir a periodicidade dos manejos físicos. É possível observar um aumento de interesse no uso dos sensores, que pode ser explicado pelo avanço da Internet das Coisas (IoT), seja na cidade ou no campo rural, criando a chamada apicultura de precisão. Dentre outras grandezas sensoriadas podemos destacar: temperatura, umidade, concentração de dióxido de carbono (CO_2) e oxigênio dentro das colmeias, massa da colmeia, padrões de imagem e intensidade de som emitido. É possível também a captura de grandezas físicas do ambiente onde a colméia se encontra, como: temperatura, umidade, vento e chuva. Para abordar os problemas descritos de determinação da saúde da colônia como um problema de Sistemas de Informação, a seguinte questão de pesquisa foi formulada:

"É possível identificar estados de saúde de colônias de abelhas *Apis mellifera* através de dados de sensores e dados do HCC? "

Para resolver essa questão, introduzimos uma abordagem de classificação usando dados de sensores internos e externos à colmeia e dados de inspeções para treino, teste e validação dos algoritmos. Dessa forma, com algoritmos de Aprendizado de Máquina treinados seria possível prover fenômenos tais como perda da rainha, abandono da colmeia e enxameação da colônia. Diante do exposto, este trabalho visa a predição do estado de saúde de colônias de abelhas *Apis mellifera*. Aqui serão usados the algorithms Naive Bayes (NB), k-Nearest Neighbors (kNN), Random Forest (RF) and Neural networks (NN) para prever tal estado baseado nos níveis de saúde reportados pelo HCC.

Trabalhos Relacionados

Reconhecer o estado de saúde de uma colônia de abelhas é de fundamental importância para o apicultor, pois de posse dessa informação ele é capaz de utilizar suas colônias de maneira mais produtiva no serviço

de polinização e produtos da colmeia, além de poder realizar o manejo de forma mais eficiente. De tal sorte que, a definição de procedimentos padronizados é vital para um reconhecimento confiável (EFSA Panel on Animal Health and Welfare 2016) e para a posterior análise dos dados coletados.

A técnicas de aprendizado de máquina não supervisionado agrupamento tem sido usada com o objetivo de identificar alguns fenômenos específicos de colônias de *Apis mellifera*, como a termorregulação (Kridi et al. 2014) e padrões sazonais (Maciel et al. 2018). Além disso, o aprendizado supervisionada tem sido usado para detectar anomalias de comportamento (Carvalho et al. 2018), presença de rainha (Robles-Guerrero et al. 2019), produção de crias e enxame (Kviesis and Zacepins 2016). No entanto, nenhum dos trabalhos citados anteriormente prevê o estado geral de saúde da colônia, utilizando dados do ambiente onde as colmeias estão e dados de inspeções realizadas pelo apicultor. Até onde sabemos, nosso estudo é o primeiro que detecta e caracteriza os estados de saúde de uma colônia de *Apis mellifera* através da mineração de dados usando os dados do microclima interno da colmeia (temperatura e umidade internas), produtividade (peso), clima externo e inspeção. A Tabela 1 resume os principais aspectos dos trabalhos relacionados apresentados.

	IoT (dados internos)	IoT (dados internos) e dados externos	IoT (dados internos), dados externos e inspeções
Kridi et al. 2014	✔	**x**	**x**
Maciel et al. 2018	✔	✔	**x**
Carvalho et al. 2018	✔	**x**	**x**
Robles-Guerrero et al. 2019	✔	**x**	**x**
Kviesis and Zacepins 2016	✔	**x**	**x**
Este trabalho	✔	✔	✔

Tabela 1. Resumo dos trabalhos relacionados

Materiais e Métodos

Esta seção descreve os aspectos metodológicos da pesquisa realizada, bem como as ferramentas utilizadas, a coleta de dados, o pré-processamento, as estratégias de aprendizagem.

Conjuntos de dados

A colméia utilizada nesta pesquisa pertence ao projeto "Bayer Bee Care Center" (BBCC) e estava localizada em Durham, Carolina do Norte, EUA. A colméia utilizada neste estudo foi a HT-101 (Lat: 35.93º N, Lon: 78.85º W, Elev: 404 pés) que possui uma colônia de abelhas *Apis mellifera*. As informações meteorológicas da região da colmeia foram obtidas a partir de uma estação meteorológica do Serviço Nacional de Meteorologia dos EUA, em inglês, National Weather Service (NWS[1]) localizada no aeroporto de Durham[2] (Lat: 35.89°N Lon: 78.78°W Elev: 394 pés).

Os dados de sensores internos à colmeia foram tomados com um amostragem diária, através da média dos valores capturados a cada hora. As seguintes variáveis físicas foram obtidas do interior da colméia:

- temperatura do aglomerado de abelhas, unidade: º Fahrenheit, temperatura obtida com um sensor colocado no centro da colônia;
- umidade do aglomerado de abelhas, umidade %, obtida com um sensor colocado no centro da colônia;
- temperatura de colméia, unidade: º Fahrenheit, temperatura dentro da caixa de madeira onde está a colônia;

[1] https://www.weather.gov/
[2] https://www.weather.gov/rah/

- umidade da colméia, umidade %, dentro da caixa de madeira onde está a colônia, e a,
- massa da colméia, unidade: libra, peso do conjunto colméia + colônia obtido através de uma balança digital.

Para amostragem do sensores externos foi tomada uma média diária de cada grandeza, embora a amostragem do sensor ocorresse a cada 5 minutos. As grandezas externas obtidas foram:

- temperatura externa, unidade: º Celsius, a temperatura do ar.
- ponto de orvalho, unidade: º Celsius, a temperatura em que um conteúdo de vapor de água e uma determinada parcela de ar (ao ser resfriada e a pressão constante) para que ocorra a saturação.
- pressão, unidade: hectopascais, a pressão do ar em relação ao nível médio do mar ou Mean Sea Level (MSL).
- velocidade do vento, unidade: m/s, a taxa de deslocamento horizontal do ar além de um ponto fixo.
- precipitação 1hr, unidade: milímetros, a espessura da precipitação líquida que é medida ao longo de um período de acumulação de uma hora.

As inspeções "in loco" foram realizadas 1 vez por semana. O HCC utilizado neste estudo foi o proposto por (Rogers 2017). Nesse HCC são observadas seis grandes informações internas da colméias, são elas:

- a presença de todas as fases de cria e instares;
- quantidade de abelhas adultas e estrutura de idades para cuidar das crias e desempenhar as tarefas da colônia;
- uma rainha jovem e produtiva;
- disponibilidade de forragem nutritiva e reservas de alimento dentro e fora da colmeia;
- a presença de estressores aparentes que impactam a subsistência da colônia; e
- a quantidade de espaço sanitário e de defesa dentro e ao redor da colmeia para a colocação de ovos e crescimento da colônia;

Uma validação dessas características foi feita por Jacobs et al (2017). Para cálculo do estado de saúde da colônia, cada característica do HCC recebeu um valor, dependendo do que foi observado, e a partir desses valores, a colônia foi categorizada em relação a sua saúde. Existem quatro categorias possíveis: 0 - doente; 1 - fraca; 2 - menos saudável; 3 - saudável. A Tabela 2 apresenta um resumo da amostragem realizada.

	Quantidade de amostras	Período de observação
Sensores Internos	180	de 4 de Abril a 30 de Setembro de 2017
Sensores Externos	180	de 4 de Abril a 30 de Setembro de 2017
Inspeções	24	de 4 de Abril a 25 de Setembro de 2017

Tabela 2. Resumo da amostragem realizada

Pré-processamento

O pré-processamento é uma etapa importante na análise de dados pois explora e analisa os dados para compreender melhor o que eles representam. Muitas vezes, podem ocorrer incoerências, tais como: o aparecimento de outliers ou dados redundantes que atrapalham o processamento e podem originar em um modelo preditivo sem qualidade.

Inicialmente, na etapa de pré-processamento, foi realizada a análise exploratória dos dados através das estatísticas básicas e o gráficos de dispersão. Em seguida, realizamos a detecção e retirada de outliers e, finalmente, a padronização do conjunto de dados pela média e desvio padrão.

Análise exploratória dos dados

Inicialmente, o primeiro passo para se familiarizar com os dados é realizar uma análise inicial com as técnicas básicas de estatísticas. Ou seja, deve-se calcular a média, desvio padrão, análise inter-quartis e obliquidade. Além de ser importante plotar o histograma para ver a distribuição dos dados.

A análise exploratória de dados emprega grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na sua estrutura, descobrir variáveis importantes em suas tendências, detectar comportamentos anômalos do fenômeno, testar se são válidas as hipóteses assumidas, escolher modelos e determinar o número ótimo de variáveis.

Gráfico de dispersão

A Figura 1 apresenta o gráfico de dispersão das variáveis internas da colméia, do peso e da temperatura externa. Como é possível observar, no primeiro mês de observação, Abril, a colônia apresenta um bom controle térmico, uma perda repentina de peso, seguida de um ganho gradual de peso. No segundo mês, Maio, é possível observar novamente uma perda acentuada de peso e um controle menor da temperatura interna pela colônia. De acordo com a planilha de inspeção, no dia 08/05/2017 a colônia não apresentava a presença de todas as fases de cria e instares, nem adultos suficientes, nem disponibilidade de recursos florais para forrageamento, sem estressores e pouco espaço dentro da colméia. Nessa inspeção a colônia foi classificada como fraca. No começo do mês de junho a colônia, aparentemente, volta a ter um controle da temperatura, contudo, em meados do mês observa-se uma queda brusca da temperatura, nesse mesmo período observa-se também uma diminuição do peso, contudo a colônia foi classificada como saudável. A colônia só retoma o controle homeotérmico em meados de Julho. Em relação ao peso, a colméia apresenta um ganho repentino de massa em 22/06/2017 e, na sequência, um ganho progressivo.

Figura 1. Dispersão variáveis entre os dias 04 de abril e 31 de setembro de 2017.

Estatísticas básicas

Na Figura 2 é possível observar às estatísticas básicas das variáveis internas da colméia, além do peso e da temperatura externa. De maneira geral, o desvio padrão de todas as amostras pode ser considerado baixo, tomando com referência suas respectivas médias amostrais. É possível observar também uma grande semelhança entre os valores da temperatura na área de cria e da colméia e entre as umidades. Indicando forte correlação entre esses pares de *features*.

	Temp-Btm(F)	Temp-Brood(F)	Temp-Hive(F)	BRH(%)	HRH(%)	Weight(lbs)
count	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000
mean	76.143333	87.266667	84.062778	58.772222	56.526111	47.710556
std	6.921548	7.086489	9.161134	6.393345	7.795260	7.219639
min	54.800000	71.900000	71.300000	27.000000	25.100000	27.900000
25%	72.650000	84.525000	72.700000	55.675000	53.400000	43.575000
50%	77.300000	89.200000	88.050000	60.100000	57.850000	46.100000
75%	81.025000	92.700000	92.700000	62.900000	61.850000	51.725000
max	88.100000	95.400000	95.100000	70.600000	69.800000	64.900000

Figura 2. Resumo das estatísticas básicas.

Detecção e remoção de anomalias

As anomalias são dados cujos valores são muito diferentes em relação aos demais ou que estão fora dos intervalos aceitáveis do conjunto de dados. As anomalias podem enviesar o resultado e consequentemente fazer com que ele apresente distorções. A detecção de anomalias pode ser realizada por meio do método de Tukey.

O método de Tukey define um outlier como aqueles valores do conjunto de dados que estão distantes do ponto central, que é mediana. A distância máxima até o centro dos dados que serão permitidos é chamada de parâmetro de limpeza, caso um dado esteja fora desse alcance ele será entendido como uma anomalia. Se o parâmetro de limpeza for muito grande, o teste se tornará menos sensível às anomalias. Pelo contrário, se for muito pequeno, muitos valores serão detectados como outliers. Esse método funciona bem quando tem-se valores de anomalias extremas e podem ser facilmente detectadas. Os limites são calculados por (1), em que Q_1 e Q_3 são, respectivamente, os primeiro e terceiro quartis de um atributo do conjunto de dados.

$$[Q_1 - 1.5 \times (Q_3 - Q_1), Q_3 + 1.5 \times (Q_3 - Q_1)] \quad (1)$$

Redimensionamento dos dados (padronização)

Após a remoção de outliers, procedemos para a fase de padronização dos dados. O redimensionamento de dados lida com parâmetros de diferentes unidades e escalas. Principalmente quando deve-se comparar valores e para isso eles precisam ter a mesma escala para acarretar bons resultados. Cada atributo "*x*" do dataset possui se valor normalizado "x_{new}" padronizado pela transformação da média e da variância ou transformação z-score (2).

A padronização redimensiona os dados para ter uma média (μ) de 0 e desvio padrão (σ) de 1 (variação unitária). Para padronizar os dados faz-se a centralização, em que o valor médio do preditor é subtraído de todos os valores. Como resultado dessa centralização, o preditor tem uma média zero. E também faz-se o escalonamento em que cada valor da variável preditora é dividido por seu desvio padrão. Escalar os dados é forçar os valores a terem um desvio padrão comum de um.

$$X_{new} = \frac{x - \mu}{\sigma} \quad (2)$$

Estratégias de Aprendizado de Máquina

A criação do modelo de classificação foi realizada por meio de mineração de dados, utilizando as seguintes etapas: (i) criação, treino e teste dos modelos de classificação (aprendizado supervisionado) e (ii) validação da classificação através de análises de especialistas em apicultura e com apoio das planilhas de inspeções (aprendizado semi-supervisionado).

Algoritmos de Classificação

A tarefa de classificação consiste em analisar várias amostras de um determinado conjunto de dados rotulados e aprender com ele um padrão para atribuir às novas amostras um rótulo entre os existentes, de acordo com sua similaridade com as amostras da classe correspondente a esse rótulo. Por ter esses rótulos baseados, é considerada uma técnica de aprendizado supervisionada. Os algoritmos utilizados nesta etapa foram Naive Bayes (NB), k-Nearest Neighbors (kNN), Random Forest (RF) e Neural Networks (NN).

- Naive Bayes (NB), baseia-se no Teorema de Bayes para gerar as predições para cada observação, classificando uma amostra em um grupo que possua a maior probabilidade de tê-lo de acordo com os atributos.
- k-vizinhos mais próximos ou k-Nearest Neighbors (kNN), baseia-se em receber um conjunto de treino, aprender com este conjunto, validar a aprendizagem com um conjunto de teste e, ao receber novas observações, classificá-las de acordo com as conhecidas. Cada nova observação tem sua distância

calculada para cada observação já conhecida. A classificação é então realizada de acordo com o maior número de k vizinhos mais próximos pertencentes à mesma classe. Uma aproximação inicial para o valor de k pode ser dada pela raiz quadrada do número de observações presentes no conjunto de dados.

- Florestas Randômicas ou Random Forest (RF), baseia-se em árvores de decisão para gerar suas classificações. Árvores de decisão são estruturas que baseiam-se em regras de decisão que se ramificam em possibilidades e criam um "caminho". No final do caminho está a classificação atribuída à entrada.
- Redes Neurais ou Neural networks (NN), baseia-se em uma metáfora do comportamento do cérebro. São formadas por unidades de processamento simples denominados neurônios que são responsáveis pelo cálculo de determinadas funções matemáticas. Os neurônios ficam organizados em uma ou mais camadas e interligados por conexões.

Métricas de avaliação dos algoritmos de classificação

Para avaliar os modelos propostos, foram utilizadas 4 métricas: (i) Precisão da Classificação ou Classification Accuracy (CA), (ii) Precisão ou Precision, (iii) Sensibilidade ou Recall e (iv) F1-score, definidas como:

- Classification Accuracy (CA) é precisão do modelo de classificação, pode ser calculada através da expressão (3);

$$CA = \frac{vp + vn}{vp + vn + fp + fn} \quad (3)$$

- Precision expressa a proporção de amostras corretamente classificadas, considerando o conjunto de todas as amostras classificadas (correta e incorretamente). Valores próximos de 1 significam que o algoritmo retorna mais resultados relevantes do que irrelevante. Pode ser calculada através da expressão (4);

$$Precision = \frac{vp}{vp + fp} \quad (4)$$

- Recall or sensibilidade explica com que eficácia o classificador identifica previsões positivas. Ou seja, a capacidade do modelo de identificar corretamente quais amostras pertencem a uma classe; Calculada por (5);

$$Recall = \frac{vp}{vp + fn} \quad (5)$$

- F1-score é uma maneira de equilibrar a Precisão e a Sensibilidade, mas sem sofrer com o problema que a precisão sofre quando há um grande desequilíbrio de classes. É a média harmônica entre precisão e sensibilidade (6);

$$F1\text{-}score = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

Onde vp = verdadeiro positivo, vn = verdadeiro negativo, fp = falso positivo e fn = falso negativo.

Configuração do Experimento e Resultados

Uma vez que os dados sensoriados não possuem indicação de classes (labels), então as classes foram obtidas através da planilha de inspeção. As classes foram atribuídas considerando que os valores permaneciam os mesmos entre uma inspeção e outra até que o estado da colônia mudasse. Essa atribuição resultou na distribuição observada na Tabela 3. Na sequência, para todos os algoritmos, foi realizada a separação do dataset nos conjuntos de treino validação e teste.

Classe	0	1	2	3
Número de amostras	2	6	97	75

Tabela 3. Quantidades de Amostras por Classe de Padrões de Saúde

O conjunto de treinamento foi usado para criar um modelo inicial para cada um dos algoritmos, com os hiperparâmetros inicialmente escolhidos, as previsões deste modelo são comparadas com os resultados

originais já conhecidos do conjunto de treinamento e então otimizados no conjunto de validação. A proporção de dados em cada um desses conjuntos é a seguinte: 60% para treino, 20% para validação e 20% para teste.

Para o Naive Bayes, por se tratar de um modelo de kernel Gaussiano, não há hiperparâmetros para serem escolhidos. Assim, um modelo inicial foi criado e executado 50 vezes. Após 50 execuções, a média e o desvio padrão da acurácia foram calculados.
Para o k-NN, um modelo inicial foi criado escolhendo o hiperparâmetro k = p, onde p é o número de diferentes preditores (features). No conjunto de validação, uma validação cruzada de 10 dobras (10-fold cross-validation) foi executada no valor de k, chegando então ao valor ótimo de k = 1.
Para as Florestas Randômicas, de acordo Oshiro, Perez, and Baranauskas (2012), o hiperparâmetro *n_estimators* (o número inicial de árvores que o algoritmo constrói antes de tomar a votação máxima ou tirar as médias das previsões) foi definido como 96. Os autores afirmam que o número inicial de árvores pode ser definido entre 64 e 128. Para obter o valor ideal das árvores, foi realizada sobre o número de árvores uma validação cruzada de dez dobras (*10-fold cross-validation*). O melhor número de árvores obtidas foi de 76.
Para as Redes Neurais, foi utilizado uma rede MLP (*Multilayer Perceptron*). A implementação utilizada do MLP treina iterativamente, uma vez que, a cada iteração, as derivadas parciais da função de perda são calculadas em relação aos parâmetros do modelo para atualizar os próprios parâmetros. A função de ativação das camadas ocultas utilizada foi a Rectified Linear Unit (ReLU), o *solver* para a otimização dos pesos foi o Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS). Foi utilizada a arquitetura 1:5:2:1, ou seja, com 1 camada de entrada, 2 camadas ocultas (com 5 e 2 neurônios) e 1 camada de saída.

Resultados

Em seguida, após a definição dos valores dos hiperparâmetros, foi possível executar a fase de treinamento dos algoritmos de classificação. Os resultados obtidos podem ser vistos nas Tabelas 4, 5, 6 e 7. A diagonal principal mostra as classes corretamente classificadas. As colunas indicam os valores preditos pelos algoritmos e as linhas os valores reais, indicados pela planilha de inspeção.

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**5**	1	0
C2	0	2	**80**	15
C3	0	0	34	**41**

Tabela 4. Matriz de Confusão do NB

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**4**	1	1
C2	0	0	**78**	19
C3	0	0	19	**56**

Tabela 5. Matriz de Confusão do kNN

	C0	C1	C2	C3
C0	**2**	0	0	0
C1	0	**4**	1	1
C2	0	0	**75**	22
C3	0	0	18	**55**

Tabela 6. Matriz de Confusão do RF

	C0	C1	C2	C3
C0	**0**	0	0	2
C1	0	**0**	0	6
C2	0	0	**78**	19
C3	0	0	14	**61**

Tabela 7. Matriz de Confusão do NN

A Tabela 8 apresenta um resumo do cálculo das métricas para cada um dos quatro algoritmos. Como é possível observar, o algoritmo kNN obteve o melhor desempenho com uma acurácia que chegou a 80%. Ou seja, 80% das classes preditas correspondiam com as classes indicadas HCC pelo apicultor.

	CA	Precision	Recall	F1-score

NB	60.11%	0.59	0.72	0.65
kNN	**80.04%**	**0.80**	**0.80**	**0.79**
RF	78.20%	0.78	0.78	0.78
NN	77.22%	0.77	0.75	0.76

Tabela 8. Métricas de avaliação da acurácia dos algoritmos de classificação.

Para os algoritmos kNN, RF e NN, que obtiveram acurácias próximas de 80% é possível observar nas suas respectivas matrizes de confusão que a maioria das predições erradas ocorreram nas classes 2 e 3. Isso pode estar relacionado com uma distinção não muito clara entre essas duas classes no HCC. O que sugere que, na verdade pode se tratar de apenas um única classe ou de 3 ou mais classes. O mesmo pode ser observado no algoritmos NB. Para os algoritmos kNN, que obtiveram acurácia de 80% (Figura 3), vale destacar que esse algoritmo passou pelo processo de cross-validation para refinar o processo de escolha da melhor combinação os hiperparâmetros.

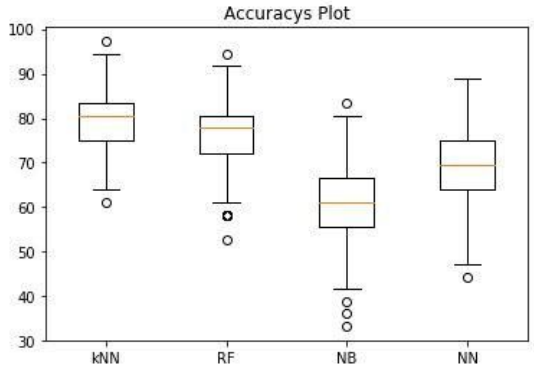

Figura 3. Boxplot com a taxa de sucesso dos algoritmos de classificação.

Conclusão

Nesse artigo, apresentamos um uma abordagem baseada em aprendizado de máquina para a predição de estados de saúde de colônias de abelhas *Apis mellifera*. Os modelos preditivos gerados utilizam dados de sensores internos e externos a uma colméia que foi monitorada por um período de 6 meses. Foram gerados 4 modelo com os algoritmos de classificação Naive Bayes, k-Nearest Neighbors, Random Forest e Neural Networks. Com taxas de acertos que chegaram a 80% com algoritmo kNN, obtendo assim a melhor acurácia.

Portanto, é possível responder a questão de pesquisa proposta de maneira positiva. Ou seja, através de dados de sensores implantados dentro de uma colméia e de dados de sensores externos (meio ambiente) é possível determinar com uma alta precisão o estado de saúde de uma colônia de abelhas *Apis mellifera*, reduzindo assim a necessidade de inspeções invasivas. Dessa forma, a solução proposta pode auxiliar de modo decisivo o apicultor, evitando perdas de colônias e auxiliando no manejo correto de suas colméias.

Como estudos futuros a curto e médio prazos, planejamos aplicar nosso método em conjuntos de dados adicionais, por exemplo, dados de colônias de abelhas africanizadas (*Apis mellifera*) no Brasil. Para isso, estamos desenvolvendo um sistema de monitoramento remoto. Outro possível trabalho futuro é aplicar essa metodologia em um conjunto maior datasets de colônias nos EUA para prever estados indesejáveis. Pretende-se também validar como mais dados o HCC utilizado, a fim de provar se as 6 características observadas na colônia durante a inspeção são suficientes para determinação de maneira geral do estado de saúde de uma colônia. Pretende-se ainda verificar outras arquiteturas de Redes Neurais.

Este estudo ressalta a importância da ciência de dados e como este tipo de metodologias pode ser construído e integrado em uma operação para ajudar as abelhas e serviços de polinização que elas

fornecem. Assim, podemos monitorar nossas colmeias remotamente e com mais precisão. Ao longo do tempo, à medida que aprendermos a interpretar os sensores e usá-los para monitorar mudanças de estado nas operações, isso nos ajudará a progredir de colmeias inertes para colmeias inteligentes e, eventualmente, colmeias geniais, conforme descrito por Cazier (2018).

Agradecimentos

Esse estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES) - código de financiamento 001. Danielo G. Gomes e Breno M. Freitas agradecem o suporte financeiro do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brasil) processos #302934/2010-3, #311878/2016-4 e #432585/2016-8.

REFERÊNCIAS

Cazier, J.A., 2018, Peering into the future a path to the genius hive - bee culture. URL: https://www.beeculture.com/peering-into-the-future-a-path-to-the-genius-hive/. 669 pages 44-46. Accessed 14 Fevereiro. 2019.

de Carvalho, H. V. F., Carvalho, E. C., Arruda, H., Imperatriz-Fonseca, V., de Souza,P., and Pessin, G. (2018). Detecção de anomalias em comportamento de abelhas utilizando redes neurais recorrentes. In 9o Workshop de Computação Aplicada a Gest ˜aodo Meio Ambiente e Recursos Naturais (WCAMA CSBC 2018), volume 9, Porto Alegre, RS, Brasil. SBC.

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

Freitas, B.M., Sousa, R.M., Bomfim, I.G.A., 2007. Absconding and and migratory behaviors of feral Africanized honey bee (Apis mellifera L.) colonies in NE Brazil. Acta Scient. Biol. Sci. 29, 381–385.

Jacobs, M., Cazier, Joseph A., Wilkes, James T., Rogers, Richard, Hassler, Edgar E. Hassler, 2017, "Building a Business Analytics Platform for Enhancing Commercial Beekeepers' Performance: Descriptive Validation of a Data Framework for Widespread Adoption By Citizen Scientists". AMCIS.

Kridi, D.S., Carvalho, C.G.N.D., Gomes, D.G., 2014. A predictive algorithm for mitigate swarming bees through proactive monitoring via wireless sensor networks. In: Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks – PE-WASUN'14. ACM Press, New York, New York, USA, pp. 41–47. http://dx.doi.org/10.1145/2653481.2653482 http://dl.acm.org/citation.cfm?doid=2653481.2653482.

Kviesis, A. and Zacepins, A. (2016). Application of neural networks for honey bee colony state identification. In2016 17th International Carpathian Control Conference(ICCC), pages 413–417.

Maciel, F. A., Braga, A. R., da Silva, T., Freitas, B., and Gomes, D. (2018). Reconhecimento de padr˜oes sazonais em colônias de abelhas apis mellifera via clusterização.Revista Brasileira de Computação Aplicada, 10(3):74–88.

Ollerton, J., Winfree, R., Tarrant, S., 2011. "How many flowering plants are pollinated by animals?" in Oikos 120, 321–326. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0706.2010.18644.x, doi:10.1111/j.1600-0706.2010.18644.x

Potts, S. G., Roberts, S. P. M., Dean, R., Marris, G., Brown, M., Jones, R., and Settele, J., 2009. Declines of managed honey bees and beekeepers in europe. Journal of Apicultural Research, (49):15–22.

Robles-Guerrero, A., Saucedo-Anaya, T., Gonz ´alez-Ram ´ırez, E., and la Rosa-Vargas, J.I. D. (2019). 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, 159:69 – 74.

Tukey, J.W., 1977. Exploratory Data Analysis. volume 2. Addison-Wesley Publishing Company.