Revista Brasileira de Agrotecnologia
V. 11, Nº 1, p. 08-19, ANO. 2021
Garanhuns, PE, Grupo Verde de Agroecologia e Abelhas
https://www.gvaa.com.br/revista/index.php/REBAGRO/index
DOI: 10.18378/rebagro.v11i1.8679

ARTIGO CIENTÍFICO

Um sistema baseado em machine learning para apoio à decisão no gerenciamento de produção apícola

A Machine Learning-Based System for Decision Support in Beekeeping Production Management

Weskley Damasceno Silva[1], Silas Santiago Lopes Pereira[2], Daniel Santiago Pereira[3*], Michell Olívio Xavier da Costa[4]

RESUMO: O setor apícola tem ganhado grandes proporções nos últimos tempos em termos de produção e comercialização de produtos, como o mel e seus derivados. O Brasil, apesar de ter acompanhado esse crescimento e possuir boas características para o desenvolvimento da apicultura, ainda sofre com a limitação no uso de ferramentas tecnológicas, o que afeta diretamente os níveis de produção. Este artigo propõe o desenvolvimento de uma ferramenta tecnológica que auxilie o apicultor no gerenciamento eficiente da produção apícola e na tomada de decisão a partir de modelos preditivos baseados em *Machine Learning* (ML) e integrados a um sistema web. Para tanto, foram utilizados diferentes algoritmos de ML para predição de produção de mel, tais como a Regressão Linear Múltipla, *Decision Tree*, *Random Forest*, *Multilayer Perceptron (MLP)* e Support Vector Regression (SVR). Os modelos gerados foram avaliados com base no coeficiente de determinação (R2 ou Score) e o cálculo de erro das predições utilizando a *Root Mean Squared Error* (RMSE). Os resultados desta pesquisa contam com um sistema web em desenvolvimento e resultados dos experimentos realizados, que mostram uma melhor performance da técnica MLP com Score de 0.98 e RMSE de 711196 libras.

Palavras-chave: Apicultura, Aprendizado de Máquina, Regressão, Modelos Preditivos, Sistemas de Informação.

ABSTRACT: The beekeeping sector has recently gained large proportions in terms of production and marketing of honey products. In the Brazilian scenario, despite having favorable conditions for beekeeping practice, the beekeeping area suffers from the limitation in the use of technological tools, which directly affects production levels. This paper proposes the development of an integrated web system to support the beekeeper in the efficient management of beekeeping production and decision making by the use of predictive models based on Machine Learning (ML). For this purpose, a comparative analysis of different ML algorithms was performed to predict honey production, such as Multiple Linear Regression, Decision Tree, Random Forest, Multilayer Perceptron (MLP), and Support Vector Regression (SVR). The generated models were evaluated based on the coefficient of determination (R2 score) and error calculation of the predictions using the Root Mean Squared Error (RMSE). Research results comprise a web system under development and experimental evaluation of regression methods. Results show the MLP algorithm obtaining better performance when compared to the other machine learning regression methods, with R2 score equal to 0.98 and RMSE equal to 711196 pounds.

Key words: Beekeeping, Machine Learning, Regression, Predictive Models, Information Systems.

1.INTRODUÇÃO

Com a crescente disseminação da informação, um grande volume de dados é gerado e armazenado a todo momento. Com o auxílio da Inteligência Artificial (IA), sistemas baseados em *Machine Learning* (ML), juntamente com a Ciência de Dados, surgem como alternativa na análise e integração desses dados a fim de ajudar em diferentes problemas reais. Dentre os vários problemas onde a IA tem sido inserida, é possível destacar aplicações na agropecuária, ecologia e meio ambiente, finanças e na saúde (CARVALHO et al., 2011).

A apicultura, inserida no meio agropecuário, destina-se à criação e exploração racional de abelhas, praticada pelo pequeno produtor rural ou agricultor familiar (MARANHÃO et al., 2016), gerando lucros e renda para muitas famílias no mundo.

Recebido para publicação em 09/02/2021, aprovado em 05/03/2021 e publicado em 14/03/2021.
*Autor para correspondência
[1] Bacharel em Ciência da Computação-Instituto Federal de Educação, Ciência e Tecnologia do Ceará-CE, Brasil. E-mail: weskleydamasceno@gmail.com;
[2] Mestre em Ciência da Computação, professor do curso Ciência da Computação no Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) campus Aracati. E-mail: silas@lar.ifce.edu.br;
[3] Doutor em Ciência Animal pela Universidade Federal Rural do Semi-Árido-UFERSA e pesquisador em Sistemas Sustentáveis / Apicultura Sustentável na Empresa Brasileira de Pesquisa Agropecuária, Embrapa Amazônia Oriental. E-mail: daniel.pereira@embrapa.br;
[4] Mestre em Ciência da Informação pela Universidade Federal do Rio de Janeiro e Analista A da Empresa Brasileira de Pesquisa Agropecuária e Diretor Adjunto da Sociedade dos Usuários de Informática e Telecomunicações do Pará. E-mail: michell.costa@embrapa.br.

Essa atividade só tem crescido nos últimos tempos com a produção e comercialização de vários produtos. Além do mel, o produto mais popular, existem outros produtos provenientes da abelha, como a cera, a geleia real e a própolis. Os preços desses produtos podem variar muito e dependem da qualidade. No âmbito mundial, a produção de mel tem se mantido nas últimas décadas com uma taxa de crescimento anual de 1,6% (GRANDÓN et al., 2016). O Brasil vem se mantendo em uma boa posição no *ranking* mundial já há alguns anos. De 2012 a 2018, o país passou de cerca de 33 mil toneladas para cerca de 42 mil toneladas de mel produzidas, impulsionado pelo aumento da demanda e pela valorização deste como um produto saudável (IBGE, 2018). Apesar do Brasil possuir boas características de clima e flora, propícias para o desenvolvimento do setor apícola, problemas com relação ao desenvolvimento tecnológico limitado e dificuldade na adesão de ferramentas tecnológicas por parte de apicultores e pequenos produtores rurais levam à deficiência da gestão básica de sistemas e métodos produtivos (VIDAL, 2017).

O desenvolvimento tecnológico limitado do setor apícola, contando com pouca inovação na utilização de ferramentas e métodos produtivos, afeta diretamente a produção tanto em volume como em qualidade. Isto revela uma deficiência significativa na gestão básica de sistemas produtivos, muitas vezes por falta de conhecimento ou atenção aos manejos necessários ou boas práticas para o cuidado com as colmeias. (GRANDÓN et al., 2016). Em vista disso, para uma melhor organização e gestão nas atividades decorrentes da apicultura, torna-se importante o uso de mecanismos de ordenamento, gestão e tomada de decisão.

A criação racional de abelhas contribui de diversas maneiras, tanto para o ser humano quanto para o meio ambiente. A apicultura tem ganhado grandes proporções nos últimos tempos em termos de produção de produtos como o mel, por exemplo. Apesar do Brasil possuir boas características para o desenvolvimento da área, a limitação no uso de ferramentas tecnológicas afeta diretamente nos níveis de produção.

Esta pesquisa surgiu da demanda apresentada pela Embrapa Amazônia Oriental, com foco no estado do Pará, onde pôde ser constatada a relevância de uma gestão eficiente da cadeia produtiva através do evento "Oficina de Planejamento da Rota do Mel", ocorrido em 2017. O intuito do evento esteve em avaliar também o perfil de uso de tecnologias pelos apicultores e dimensionar as necessidades tecnológicas de produção, processos produtivos e fatores que interferem na cadeia de produção (EMBRAPA, 2017). A pesquisa apresentada é composta por dois fluxos de trabalho. O primeiro corresponde à proposta de um módulo *web* para interação com o usuário e gerência de informações da cadeia apícola. Em adição a isso, o segundo fluxo conta com um módulo de inteligência através da geração de modelos preditivos, fornecendo informações úteis para a tomada de decisão.

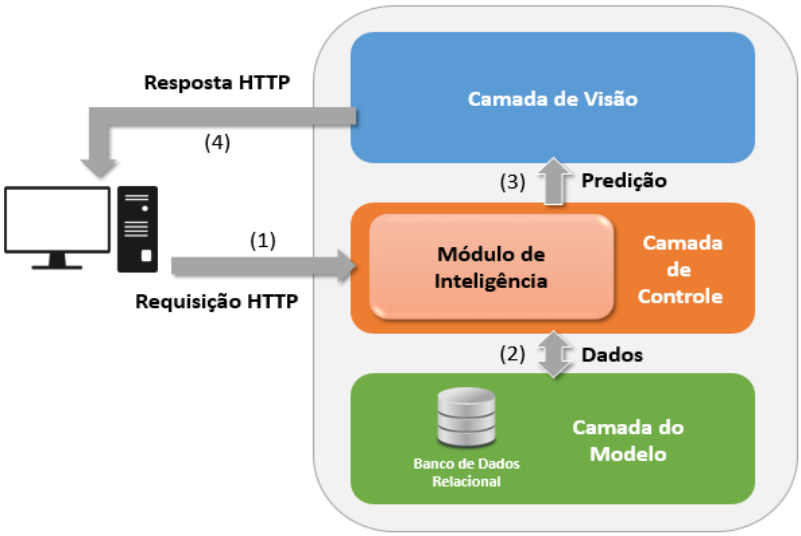
2. MATERIAL E MÉTODOS

Este trabalho consiste no desenvolvimento de um sistema *web* para gerenciamento de informações relacionadas ao contexto produtivo, além de um mecanismo de georreferenciamento que ajuda no ordenamento apícola, mostrando a disposição geográfica dos apiários da região em forma de mapa. Além disso, foram realizados experimentos e análises sobre duas bases de dados utilizando ML a fim de encontrar um modelo que será usado para a predição de produção de mel e integrado no sistema proposto.

A linguagem de programação *Python 3* foi utilizada tanto na geração dos modelos preditivos quanto no desenvolvimento do sistema *web*, com o auxílio do *microframework* Flask (http://flask.pocoo.org/). O acesso aos dados se deu por meio da conexão com o Sistema Gerenciador de Banco de Dados (SGBD) MySQL.

A arquitetura da solução é apresentada na Figura 1, onde em (1) é feita uma requisição HTTP à camada de controle, em busca de realizar uma predição para um novo exemplo cadastrado. Essa camada acessa os dados em (2) ao se comunicar com o banco de dados do sistema. O módulo de inteligência é então acionado, realizando a predição com o modelo já treinado e (3) retornando o valor solicitado à camada de visão, que se encarregará de (4) retornar uma resposta HTTP ao serviço solicitante.

Figura 1 - Arquitetura da solução.



Fonte: autores (2021).

2.2 Aquisição dos dados

O *dataset Honey Production in the USA*, disponível na plataforma Kaggle (https:// www.kaggle.com/jessicali9530/honey-production), foi utilizado com fins de verificar a possibilidade de encontrar um modelo preditivo satisfatório para o problema. O *dataset* aborda dados sobre a produção de mel nos Estados Unidos entre os anos de 1998 e 2012 além de ser composto de 8 atributos e um total de 626 amostras, descritos na Tabela 1, sendo o atributo em negrito o alvo da predição.

Tabela 1 - *Dataset honey Production in the USA* (1998-2012)

Atributo	Descrição
state	Abreviação do estado.
numcol	Número de colônias produtoras de mel.
yieldpercol	Rendimento por colmeias, dado em libras.
totalprod	Total de produção (produto do numcol x yieldpercol), dado em libras.
stocks	Ações detidas pelos produtores, em libras.
priceperlb	Preço médio por libra com base nas vendas expandidas, em dólares.
prodvalue	Valor de produção (produto do totalprod x priceperlb), em dólares.
year	Ano ao qual o dado pertence.

Fonte: autores (2021).

O segundo conjunto de dados foi obtido a partir das informações disponibilizadas pela Embrapa Amazônia Oriental, com autorização concedida pelo Ministério do Desenvolvimento Regional (MDR). Esses dados foram coletados através da aplicação de questionários no evento "Oficina de Planejamento da Rota do Mel", ocorrido em 2017, em Belém, no Pará. O questionário contou com 67 questões a respeito da identificação e caracterização de órgãos, associações, federações ou cooperativas, e as comercializações realizadas. A partir das perguntas e respostas informadas, foi possível criar um *dataset*. O *dataset* completo é composto de 231 atributos e um total de 23 exemplos, que posteriormente foi subdividido em dois *datasets* para a realização de diferentes experimentos. Na Tabela 2 podem ser conferidas as características originais de cada *dataset*.

Tabela 2 - Características originais dos *datasets*

Dataset	Nº de exemplos	Nº de atributos	Média (atributo alvo)	Desvio padrão (atributo alvo)
Kaggle	626	8	4.169.086 ,26 lb	6.883.84 6,75 lb
Embrapa	23	231	34.233,33 kg	67.783,4 3 kg

Fonte: autores (2021).

2.3 Preparação dos dados

A fim de se obter um melhor desempenho dos algoritmos, no *dataset Honey Production in the USA*, atributos categóricos foram transformados em numéricos. Já para o segundo cenário, o *dataset* foi construído obedecendo a originalidade das respostas. Para questões de múltipla escolha, foram utilizados variáveis *dummy*, atribuindo 1 ou 0 para o caso da questão estar marcada ou não, respectivamente.

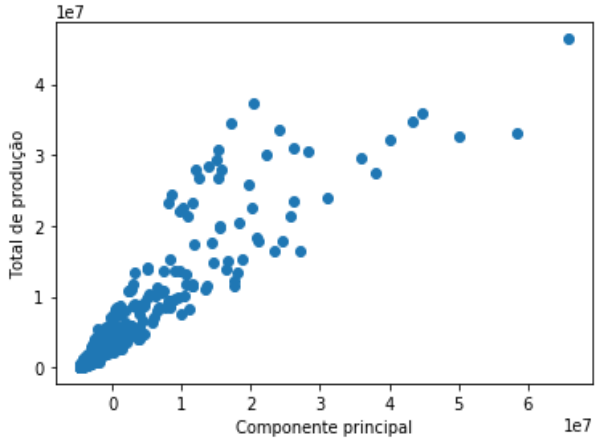
O *dataset* original conta com 231 atributos e apenas 23 amostras. Além de serem poucas amostras, alguns questionários apresentavam muitas questões sem respostas. Inicialmente, pôde-se perceber a irrelevância de alguns atributos, que resultou na remoção dos mesmos. Após tratar a inconsistência nos dados de alguns atributos e a transformação de atributos categóricos em numéricos, o \textit{dataset} passou a ter um total de 206 atributos. Com essa nova versão, foi feito também o tratamento de valores nulos. Através da classe \textit{Imputer} do \textit{scikit-learn}, foi possível pegar a média para atributos com valores contínuos. Para atributos de valores discretos, a estratégia consistiu em adicionar o valor “Não respondeu” às perguntas deixadas em branco. Após o tratamento adequado, os dados foram submetidos à etapa de processamento, onde foram submetidos a diferentes algoritmos para a geração de modelos.

2.4 Análise dos dados

Para se ter uma análise da estrutura dos dados trabalhados, foi feito o uso da técnica *Principal Component*

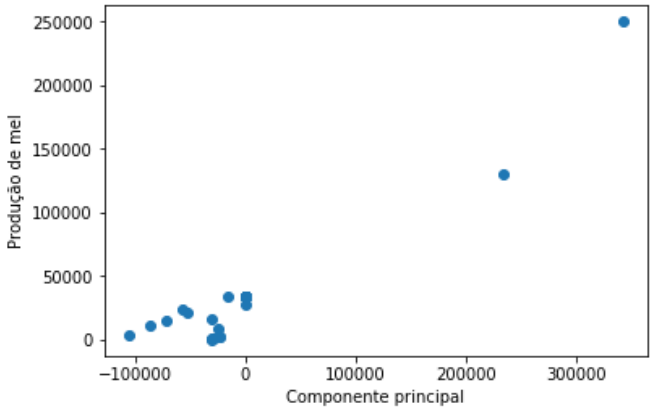
Analysis (PCA). Esta abordagem ajudou a visualizar e entender melhor o comportamento dos dados, reduzindo a dimensionalidade do vetor de entradas à 1 dimensão. Dessa forma, foi possível visualizar a dispersão dos dados a partir de um gráfico de duas dimensões, em função da produção de mel, tanto no cenário do *dataset Honey Production in the USA*, como mostra a Figura 3, como também no cenário do *dataset* da Embrapa, representado pela Figura 4.

Figura 2 - Dispersão dos dados para o *dataset Honey Production in the USA.*



Fonte: autores (2021).

Figura 3 - Dispersão dos dados para o *dataset* da Embrapa.



Fonte: autores (2021).

Com o objetivo de minimizar o problema da dimensionalidade para o segundo *dataset*, foram avaliados modelos com menores quantidades de atributos. Dessa forma, buscou-se selecionar características a partir da verificação do coeficiente de correlação de Pearson, indicando o grau de correlação de cada atributo com o atributo alvo da predição (quantidade de mel produzida). Foram considerados apenas os atributos com correlação acima de 80%. O *dataset* criado a partir dessa técnica contém 8 atributos identificados como importantes, mostrados na Tabela 3. O atributo em negrito é o atributo alvo da predição.

Tabela 3 - *Dataset* da Embrapa com os atributos de correlação acima de 80%

Atributo	Descrição
B28_5	Em geral, o empréstimo bancário realizado pela associação/federação/cooperativa/órgão tem outro objetivo que não seja o investimento, o custeio ou o capital de giro e investimento.
B39_1	As normas técnicas da série ISO são utilizadas no processo de beneficiamento do produto.
C492014MEL_P	Quantidade (kg) de mel produzida em 2014.

C492014MEL_C	Quantidade (kg) de mel comercializada em 2014.
C492015MEL_P	Quantidade (kg) de mel produzida em 2015.
C492015MEL_C	Quantidade (kg) de mel comercializada em 2015
C492016MEL_P	Quantidade (kg) de mel produzida em 2016.
C492016MEL_C	Quantidade (kg) de mel comercializada em 2016.

Fonte: autores (2021).

2.5 Modelagem e avaliação

Para um melhor processamento de alguns algoritmos de ML, os dados necessitam estar distribuídos dentro de uma mesma escala. Para isso, seguiu-se o método *StandardScaler* do *scikit-learn*. Este método consiste em subtrair a média x_{mean} dos valores de cada atributo x_i e escalonar dividindo pelo desvio padrão (DP) x_{std}. Cada valor é escalonado conforme a expressão $x_{scaled} = (x_i - x_{mean}) / x_{std}$.

Após preparar os *datasets*, na etapa 1, os atributos que irão compor cada um deles durante os experimentos são selecionados de acordo com algumas abordagens que serão mostradas mais a frente. A etapa 2 consiste em encontrar os hiperparâmetros para cada algoritmo de ML utilizado. Com esse intuito, fez-se uso da técnica *GridSearchCV* para selecionar os hiperparâmetros para cada algoritmo. Os parâmetros selecionados são submetidos a cada algoritmo através do uso da *cross-validation* com o número de partições igual a 10 (parâmetro comumente usado na literatura em experimentos de ML), utilizando todo o *dataset* para realizar o treinamento e retornando os melhores parâmetros encontrados. Os mesmos algoritmos foram avaliados para ambos *datasets* durante os experimentos, portanto, seguem a mesma *grid* de parâmetros.

2.6 Experimentos com o *dataset Honey Production in the USA*

Na realização dos experimentos envolvendo o *dataset Honey Production in the USA*, a base foi dividida utilizando a técnica *K-fold Cross-validation*, com o número de partições igual a 10. Essa técnica foi escolhida pelo fato do *dataset* possuir um tamanho razoável de amostras. Os experimentos foram executados em 30 rodadas para cada algoritmo a fim de garantir a veracidade dos resultados (JAIN, 1990). Em cada uma das rodadas, o *dataset* foi dividido em 10 subconjuntos mutuamente exclusivos e de mesmo tamanho, e executado 10 vezes, sendo sempre usado um subconjunto como teste. No treinamento, cada algoritmo de ML foi usado como regressor e, após cada rodada, foi calculada a média dos resultados de desempenho, seguindo as métricas de avaliação R^2 e o cálculo do erro através da RMSE para as partições de dados geradas aleatoriamente. Ao final das 30 execuções, foi verificada a média de todos os valores das métricas de avaliação, bem como o desvio padrão entre eles. Além disso, também foi calculado o intervalo de confiança (IC) desses resultados para um nível de confiança de 95%. A partir dos resultados dessas métricas, é possível verificar o modelo mais bem avaliado.

2.7 Experimentos com os *datasets* construídos a partir dos questionários

Para os experimentos desse cenário, foi utilizado o *dataset* completo formado a partir dos questionários. Como logo de início o resultado se mostrou péssimo, pôde-se perceber que esse *dataset* tinha uma quantidade muito grande de atributos irrelevantes que, após tratados, consistiu em outros dois *datasets* contendo apenas parcelas de atributos. O primeiro *dataset* foi gerado a partir da percepção do negócio. Logo, foram considerados arbitrariamente apenas alguns atributos que pudessem interferir diretamente no objetivo da predição, que estão presentes. Sendo assim, o *dataset* foi composto pelos atributos presentes na Tabela 4. O atributo em destaque é o alvo da predição. O *dataset* consiste de atributos com correlação acima de 80% para com o atributo alvo e pode ser verificado na Tabela 3.

Tabela 4 - *Dataset* da Embrapa com os atributos escolhidos arbitrariamente

Atributo	Descrição
B9	Anos de funcionamento que a entidade possui.
B12	Número de apicultores associados à entidade.
B13	Número aproximado de colmeias.
B34_1	Se os empreendimentos ou apicultores não utilizam programas ou aplicativos de gerenciamento.
B34_2	Se os empreendimentos ou apicultores utilizam programas ou aplicativos de gerenciamento para gerenciamento administrativo.

B34_3	Se os empreendimentos ou apicultores utilizam programas ou aplicativos de gerenciamento para contabilidade/vendas.
B34_4	Se os empreendimentos ou apicultores utilizam programas ou aplicativos de gerenciamento para produção.
B34_5	Se os empreendimentos ou apicultores utilizam programas ou aplicativos de gerenciamento para gerenciamento administrativo, contabilidade/vendas e produção.
C492015MEL_P	Quantidade (kg) de mel produzida em 2015.
C492016MEL_P	Quantidade (kg) de mel produzida em 2016.

Fonte: autores (2021).

A técnica aplicada para dividir esse *dataset* em conjuntos de treinamento e teste durante os experimentos foi a LOOCV, uma vez que a base continha um número muito reduzido de exemplos. Esta técnica consiste em dividir a base deixando sempre apenas um exemplo para formar o conjunto de teste. Os experimentos foram executados 23 vezes onde, a cada iteração, 22 exemplos eram usados para treinamento e 1 para teste. Dentro de cada iteração, cada algoritmo foi treinado usando esses conjuntos divididos e então, o resultado das métricas de avaliação R^2 *Score* e RMSE se deu pelas listas formadas dos valores reais e dos valores preditos para o conjunto de teste.

3. RESULTADOS E DISCUSSÃO

Foram realizadas de avaliações comparativas de desempenho no uso de diferentes experimentos em diferentes *datasets*. Essas avaliações puderam comprovar a eficiência de modelos preditivos, usando algoritmos de ML voltados para a tarefa de regressão, para realizar predições a respeito da produção de mel.

A etapa de análise dos resultados foi realizada a partir do uso de métricas de avaliação presentes na literatura, como forma de validar o desempenho dos modelos a partir de experimentos controlados.

3.1 Avaliação no *dataset Honey Production in the USA*

Primeiramente, são exibidos na Tabela 5 os resultados de desempenho dos modelos gerados com o *dataset Honey Production in the USA* durante a fase de estudo e a fim de verificar a possibilidade de encontrar um modelo preditivo de qualidade.

Tabela 5 - Resultados segundo as métricas de avaliação para o *dataset Honey Production in the USA*.

Algoritmo	R^2 *Score*	DP do R^2	IC do R^2	RMSE	DP da RMSE	IC da RMSE
Regressão Linear Múltipla	0,9523	0,0015982	(0,9517; 0,9529)	1.382.690 lb	6.039,84 lb	(1.380.526 lb; 1.384.849 lb)
Decision Tree	0,9624	0,0016880	(0,9618; 0,9630)	1.245.690 lb	25.653,2 lb	(1.236.513 lb; 1.254.873 lb)
Random Forest	0,9794	0,0005254	(0,9792; 0,9795)	922.693 lb	12.168,3 lb	(918.338 lb; 927.046 lb)
MLP	**0,9997**	**0,0000107**	**(0,9997; 0,9997)**	**100.812 lb**	**2.148,3 lb**	**(100.043 lb; 101.580 lb)**
SVR	0,9504	0,0016605	(0,9498; 0,9510)	1.416.260 lb	7.770,13 lb	(1.413.479 lb; 1.419.040 lb)

Os resultados seguem as métricas de avaliação R^2 *Score*, que mostra o ajuste do modelo aos dados, e o cálculo do erro a partir da métrica RMSE, que penaliza erros maiores e entrega o resultado na escala padrão dos dados. Além disso, foram verificados seus respectivos desvios padrões e intervalos de confiança para um nível de confiança de 95% para a média dos resultados.

Todos os modelos avaliados tiveram ótimos resultados e mostraram um bom comportamento para os dados trabalhados. O modelo gerado utilizando o algoritmo Regressão Linear Múltipla obteve um *Score* de cerca de 95% de acerto nas predições realizadas. Porém, o modelo teve ainda um erro de cerca de 1.382.690 libras, onde o maior valor registrado no *dataset* era de 46.410.000 e o menor de 84.000 libras. Para o modelo utilizando o algoritmo *Decision Tree*, o *Score* foi de cerca de 96% e o modelo errou cerca de 1.245.690 libras. O modelo utilizando o *Random Forest* com 50 árvores teve um *Score* por volta de 97% e a média do erro foi de 922.693 libras. O modelo utilizando SVR de *kernel* linear obteve um desempenho parecido com o modelo utilizando Regressão Linear, com 95% de *Score* e errando em média 1.416.260

libras. O melhor modelo avaliado foi o gerado a partir do uso da MLP com configuração de uma camada oculta de 10 neurônios, a função de ativação tanh e o *solver* foi o lbfgs que possui bons resultados com *datasets* pequenos. O modelo teve um ótimo *Score* de cerca de 99% e teve a menor média de erros registrada entre os modelos, com cerca de 100.812 libras.

3.2 Avaliação nos *datasets* gerados dos questionários

O primeiro *dataset* formado a partir dos questionários adquiridos, consiste de atributos escolhidos arbitrariamente de acordo com a compreensão do negócio. O *dataset* gerado pode ser conferido através da Tabela 4. Os resultados dos experimentos envolvendo o *dataset* em questão são apresentados na Tabela 6.

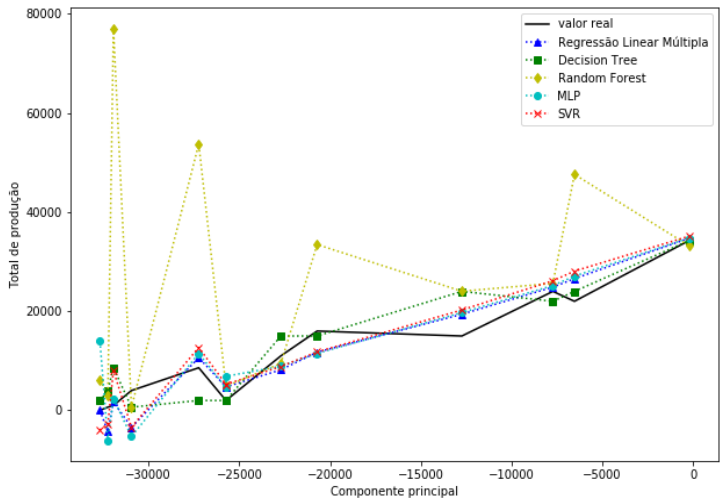
Tabela 6 - Resultados segundo as métricas de avaliação para o *dataset* com atributos escolhidos arbitrariamente.

Algoritmo	R^2 *Score*	RMSE
Regressão Linear Múltipla	**0,9916**	**4.828,8 kg**
Decision Tree	0,0464	51.640,9 kg
Random Forest	0,1406	49.023,3 kg
MLP	0,3567	42.413,9 kg
SVR	0,8634	19.545,5 kg

A partir dos resultados das métricas de avaliação R^2 *Score* e RMSE, foi possível inferir que o melhor desempenho entre os modelos, se deu através do algoritmo Regressão Linear Múltipla. Os resultados mostram um *Score* de aproximadamente 99% do modelo. O modelo também gera um erro de cerca de 4.828,8 kg de mel produzidos, que pode ser considerado pequeno se levarmos em conta que os valores do atributo alvo da predição variam entre 250.000 kg (maior valor registrado) e 15 kg (menor valor registrado). O modelo utilizando o algoritmo *Decision Tree* teve o pior resultado entre os modelos, obtendo um *Score* de 4% e 51.640,9 kg de erro, sendo o maior valor registrado. Em seguida temos o modelo usando *Random Forest* com um total de 10 árvores. O modelo teve uma leve melhora, contando com um *Score* de 14% e errando cerca de 49.023,3 kg. O próximo modelo usou MLP com a função de ativação *logistic*, uma camada oculta com 100 neurônios e novamente, se tratando de bases pequenas, o *solver* utilizado foi o lbfgs. O modelo obteve um *Score* de cerca de 35% e uma média de erro de 42.413,9 kg. O modelo utilizando o algoritmo SVR com *kernel* linear teve o segundo melhor resultado avaliado, com *Score* de 86%, mas projetando um erro ainda muito grande de 19.545,5 kg, se comparado ao melhor modelo avaliado.

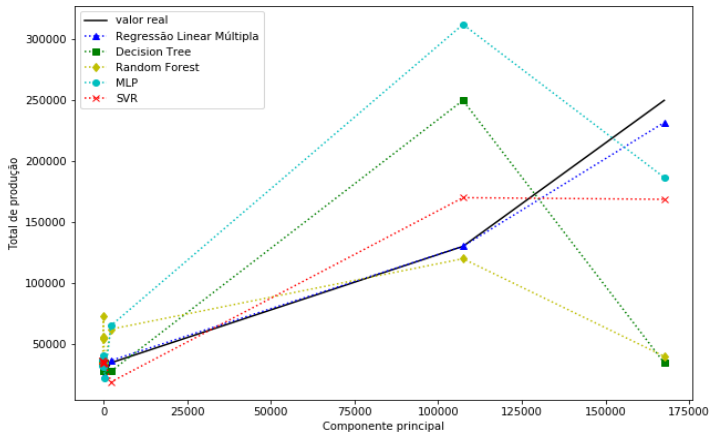
Como nesses experimentos foi utilizada LOOCV, a cada rodada um dos 23 exemplos formava o conjunto de teste. Então resolveu-se, ao final de todas as rodadas da LOOCV, montar um gráfico com todos os pontos que formaram o conjunto de teste a cada rodada. Como alguns dados ficaram expostos de maneira muito próxima, dificultando a visualização, o gráfico presente na Figura 4 contém os 12 primeiros exemplos do conjunto de dados, bem como o gráfico da Figura 5 mostra os 11 exemplos restantes.

Figura 4 - Desempenho dos 12 primeiros exemplos.



Fonte: autores (2021).

Figura 5 - Desempenho dos 11 últimos exemplos.



Fonte: autores (2021).

Através do gráfico, é possível conferir que o modelo mais bem avaliado (Regressão Linear Múltipla) teve um desempenho bem próximo do que se aproxima dos valores reais.

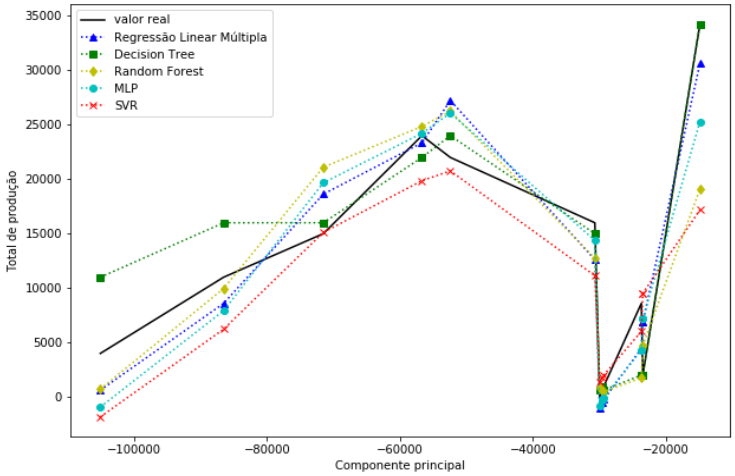
Com o intuito de melhorar ainda mais esse resultado, foram avaliados os modelos utilizados em experimentos a partir de um novo *dataset* formado com os atributos mais bem correlacionados com o atributo alvo da predição. Esta abordagem se deu com base na hipótese de que um bom conjunto de atributos são aqueles mais bem correlacionados com o atributo alvo, demonstrada por (HALL, 1999). Os atributos escolhidos foram os de correlação acima de 80%, como já citado anteriormente. Os resultados do desempenho dos modelos são apresentados na Tabela 7.

Tabela 7 - Resultados segundo as métricas de avaliação para o dataset com atributos de correlação acima de 80%

Algoritmo	R^2 *Score*	RMSE
Regressão Linear Múltipla	**0,9963**	**3.176,1 kg**
Decision Tree	0,6308	32.133,2 kg
Random Forest	0,5514	35.420 kg
MLP	0,9026	16.496,9 kg
SVR	0.9058	16.225,6 kg

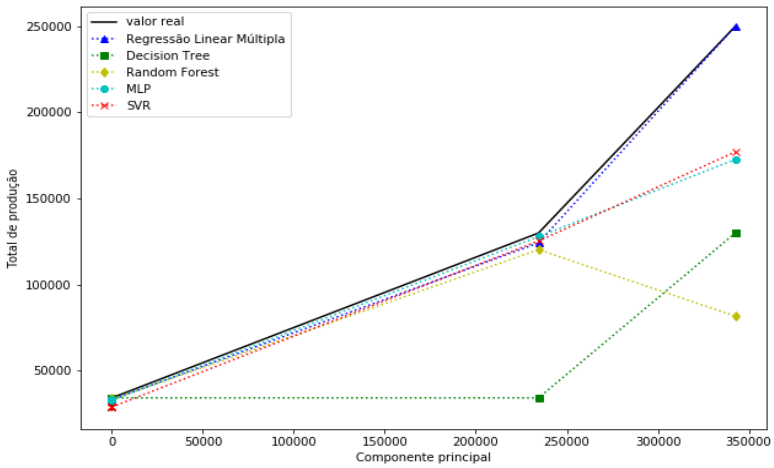
No geral, todos os modelos tiveram melhores resultados quando aplicado essa estratégia e utilizando as mesmas configurações. Ainda sim, o melhor modelo avaliado continuou sendo o que usou Regressão Linear Múltipla com um *Score* de aproximadamente 99%, gerando erros ainda menores, com RMSE de 3.176,1 kg de mel produzidos. O modelo utilizando *Decision Tree* obteve um *Score* de 63% e em média um erro de 32.133,2 kg. O modelo utilizando *Random Forest* foi um pouco pior que o anterior, obtendo um *Score* de 55% e errando cerca de 35.420 kg. Os modelos utilizando os algoritmos MLP e SVR tiveram resultados bastante parecidos, contando com um *Score* de 90% e erro de 16.496,9 kg e 16.225,6 kg, respectivamente. Ademais, como nos outros casos, a Figura 6 e a Figura 7 mostram o gráfico de desempenho dos modelos conforme os resultados da predição, dividido em duas partes para uma melhor visualização.

Figura 6 - Desempenho dos 12 primeiros exemplos.



Fonte: autores (2021).

Figura 7 - Desempenho dos 11 últimos exemplos.



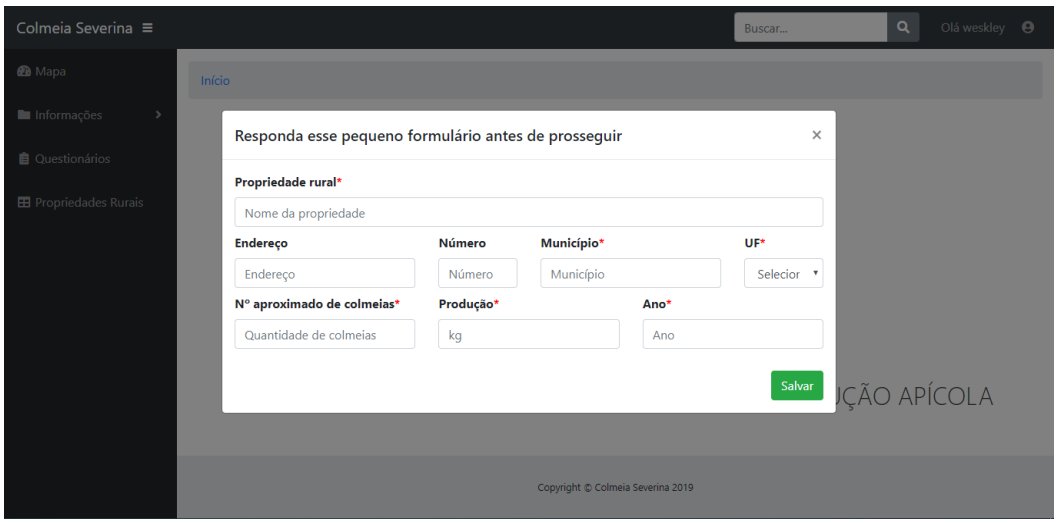
Fonte: autores (2021).

Por fim, a partir dos experimentos realizados e a avaliação dos modelos, foi possível perceber que o modelo utilizando Regressão Linear Múltipla teve o melhor ajuste ao problema abordado, obtendo os melhores resultados de desempenho na predição de produção de mel.

3.3 Sistema *Web*

O usuário, ao se cadastrar no sistema, realiza o *login* automaticamente e é direcionado a uma tela com um formulário a respeito de características sobre sua propriedade, como mostrado na Figura 8. As informações fornecidas são adicionadas ao banco de dados.

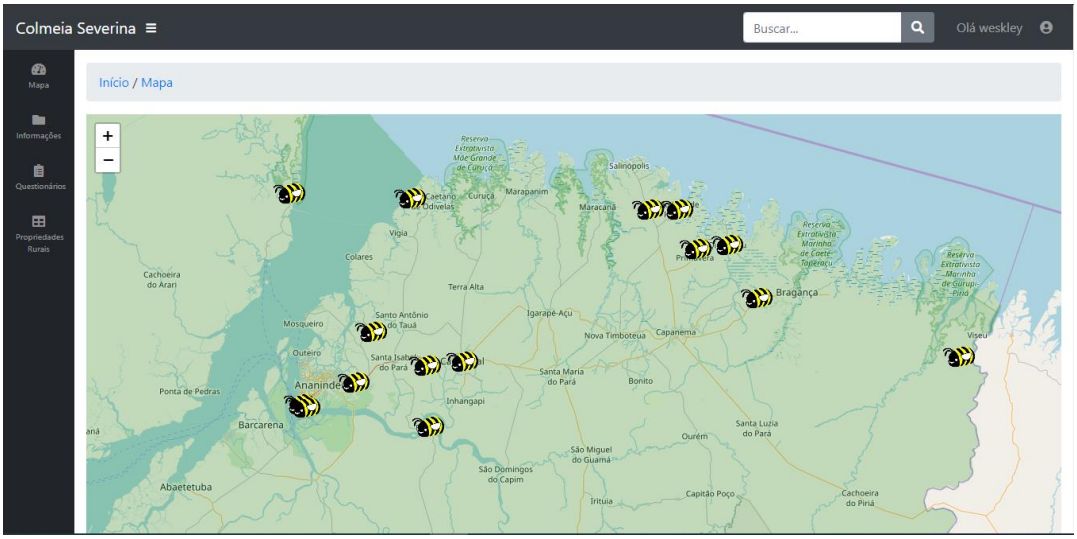
Figura 8 - Tela de formulário inicial



Fonte: autores (2021).

O usuário apicultor/gestor consegue gerenciar suas propriedades, bem como seus apiários e colmeias. Além disso, tem acesso à uma ferramenta de geolocalização de todas as propriedades cadastradas em sua região dispostas em um mapa através de marcadores, como mostra a Figura 9.

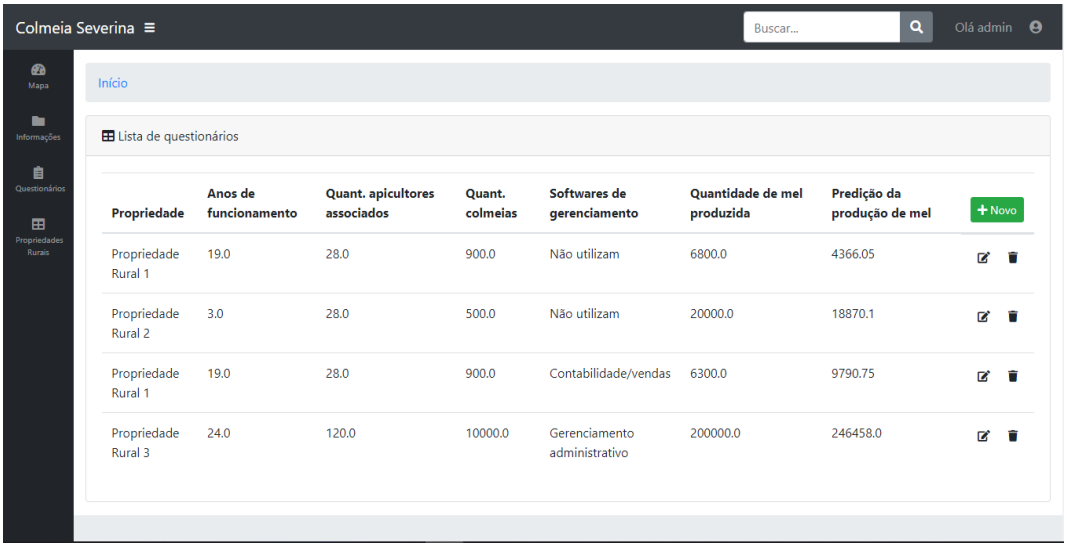
Figura 9 - Tela de georreferenciamento



Fonte: autores (2021).

Em se tratando do módulo de predição integrado ao sistema *web*, os resultados conquistados mostram uma tela onde são passadas as informações necessárias por meio de um questionário para o modelo de ML realizar a predição. Essas informações são salvas no banco de dados e é possível visualizar a predição retornada para aquela configuração de parâmetros informados pelo usuário. Essa funcionalidade pode ser vista na Figura 10.

Figura 10 - Tela contendo a realização da predição da produção de mel



Propriedade	Anos de funcionamento	Quant. apicultores associados	Quant. colmeias	Softwares de gerenciamento	Quantidade de mel produzida	Predição da produção de mel
Propriedade Rural 1	19.0	28.0	900.0	Não utilizam	6800.0	4366.05
Propriedade Rural 2	3.0	28.0	500.0	Não utilizam	20000.0	18870.1
Propriedade Rural 1	19.0	28.0	900.0	Contabilidade/vendas	6300.0	9790.75
Propriedade Rural 3	24.0	120.0	10000.0	Gerenciamento administrativo	200000.0	246458.0

Fonte: autores (2021).

4. CONCLUSÃO

Os resultados mostraram um melhor desempenho do modelo usando Regressão Linear Múltipla em ambos os *datasets* gerados a partir dos dados repassados pela Embrapa Amazônia Oriental, onde foi possível a criação de modelos preditivos. A escolha do modelo para ser posto em produção, resultou naquele que consiste de atributos selecionados arbitrariamente, que no entendimento do negócio, se mostra ser mais plausível para compreender um sistema *web* a ser utilizado pelo apicultor a fim de facilitar o fornecimento das informações. Essa escolha foi possível tendo em vista que o desempenho do modelo escolhido contou com um *Score* de cerca de 99% de acerto nas predições, apesar de contar com um erro, calculado através da RMSE, um pouco pior que o modelo com atributos de forte correlação com a produção de mel, mas ainda sim bem próximo, além de ser considerado baixo se comparado às proporções dos valores do atributo alvo da predição.

Os resultados permitiram o desenvolvimento de um sistema *web*. A linguagem de programação Python foi utilizada em ambos os módulos de trabalho (módulo *web* e módulo de inteligência), visando a simplicidade e eficiência que essa linguagem apresenta tanto para o desenvolvimento *web* quanto pelo vasto uso na ciência de dados.

Com o desenvolvimento de um sistema *web* foi possível a interação do modelo de predição com os gestores e pequenos produtores rurais a fim de conseguirem tomar decisões baseadas nos resultados das predições realizadas a partir das características de suas propriedades.

O sistema *web* mantém dados a respeito das propriedades e seus níveis de produção, constituindo uma base de dados para a realização das predições feitas pela ferramenta. Ainda sobre o módulo *web*, é possível ter acesso à geolocalização das propriedades informadas pelos usuários do sistema através da utilização de um mapa de marcadores. Com isso, é possível também tomar decisões com base na disposição geográfica dos negócios na região.

Trabalhos futuros incluem a aplicação de técnicas de seleção de atributos de forma mais exaustiva a fim de avaliar suas contribuições. Além disso, para contrapor uma das maiores dificuldades encontradas, sugere-se a aquisição de mais dados e de novas características, construindo uma base mais completa e consistente, para garantir a fidelidade dos resultados retornados pelos modelos gerados.

AGRADECIMENTOS

Ao PIAmz – Projetos Integrados da Amazônia (Fundo Amazônia / BNDES); Projeto AGROBIO - Abelhas, variedades crioulas e bioativos agroecológicos: conservação e prospecção da biodiversidade para gerar renda aos agricultores familiares na Amazônia Legal (Embrapa - SEG / Ideare 16.17.01.004.00.00); pelo financiamento a este trabalho.

Ao Governo Federal do Brasil / Ministério do Desenvolvimento Regional - Programa Rotas da Integração: Rota do Mel; pela disponibilização da base de dados da cadeia produtiva do mel do estado do Pará/Brasil.

REFERÊNCIAS

CARVALHO, ACPLF et al. Inteligência Artificial – uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

EMBRAPA (2017). SUDAM - Superintendência do Desenvolvimento da Amazônia: Oficina debate Rota do Mel na Região Norte. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/21352318/oficina-prepara-inclusao-da-amazonia-na-rota-do-mel>. Acessado em: 09-02-2021.

GRANDÓN, Natalia et al. Information system for improving local productivity and decision making in organic beekeeping. In: 2016 IEEE International Conference on Automatica (ICA-ACCA). IEEE, 2016. p. 1-7.

HALL, Mark Andrew. Correlation-based feature selection for machine learning. 1999.

IBGE (2018). Pesquisa da Pecuária Municipal. Disponível em: <https://sidra.ibge.gov.br/pesquisa/ppm/quadros/brasil/2018>. Acessado em: 05-03-2020.

JAIN, Raj. The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. John Wiley & Sons, 1990.

MARANHÃO, Patrícia Bastos de Andrade Albuquerque et al. Avaliação dos métodos de custeio na produção de mel: um estudo de caso no município de São João do Rio do Peixe. 2016.

VIDAL, Maria de Fátima. (2017). Desempenho da Apicultura Nordestina em Anos de Estiagem. Disponível em: <https://www.bnb.gov.br/documents/80223/2130269/apicultura_11_2017.pdf/>. Acessado em: 19-01-2019.