INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
IFCE CAMPUS ARACATI
COORDENADORIA DE CIÊNCIA DA COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

WESKLEY DAMASCENO SILVA

UMA FERRAMENTA PARA APOIO À DECISÃO NO GERENCIAMENTO DE PRODUÇÃO APÍCOLA BASEADA EM APRENDIZADO DE MÁQUINA

ARACATI-CE
2019

WESKLEY DAMASCENO SILVA

UMA FERRAMENTA PARA APOIO À DECISÃO NO GERENCIAMENTO DE PRODUÇÃO APÍCOLA BASEADA EM APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Msc. Silas Santiago Lopes Pereira
Coorientador: Dr. Daniel Santiago Pereira

Aracati-CE
2019

WESKLEY DAMASCENO SILVA

UMA FERRAMENTA PARA APOIO À DECISÃO NO GERENCIAMENTO DE PRODUÇÃO APÍCOLA BASEADA EM APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Aprovada em 30 de Setembro de 2019

BANCA EXAMINADORA

Prof. Msc Silas Santiago Lopes Pereira (Orientador)
IFCE

Dr. Daniel Santiago Pereira (Coorientador)
Embrapa Amazônia Oriental

Prof. Dr. Mário Wedney de Lima Moreira
IFCE

Prof. Dr. Reinaldo Bezerra Braga
IFCE

DEDICATÓRIA

À minha família que nunca mediu esforços para ajudar no que foi preciso para chegar até aqui.
E à todos os colegas e amizades conquistadas através dos momentos divididos durante esta fase da vida.

AGRADECIMENTOS

Agradeço primeiramente à Deus pelo dom da vida e por me proporcionar saber e conhecimento.
À todos os meus professores que contribuíram de alguma forma em minha formação e em meu crescimento. Agradeço principalmente ao meu orientador prof. Msc Silas Santiago, pela parceria e comprometimento na realização deste trabalho e por ser para mim, fonte de inspiração enquanto pessoa e profissional.
Agradeço também ao meu coorientador Dr. Daniel Santiago que teve imensa importância neste trabalho ajudando na disponibilização dos dados e ideias.

RESUMO

É perceptível o destaque que o setor apícola vem ganhando nos últimos anos, principalmente na produção e comercialização de produtos relacionados ao mel. No cenário brasileiro, apesar de termos condições favoráveis para a prática da apicultura, essa área sofre com a limitação no uso de ferramentas tecnológicas, o que afeta diretamente os níveis de produção. Este trabalho consiste em um sistema *Web* para apoiar o apicultor na gestão eficiente da produção e tomada de decisão da apicultura, através da utilização de modelos preditivos baseados em Aprendizado de Máquina (AM). Para este propósito, uma análise comparativa de diferentes algoritmos de AM foi realizada para predizer a produção de mel, como 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 (R^2 *Score*) e no cálculo do erro das predições através da métrica *Root Mean Squared Error* (RMSE). Os resultados da pesquisa compreendem o desenvolvimento do sistema *Web* proposto e a avaliação experimental de métodos de regressão em diferentes *datasets*. Ao utilizar um *dataset* do cenário real, o modelo que utiliza Regressão Linear Múltipla obteve melhores resultados de desempenho quando comparado com os outros modelos analisados, obtendo R^2 *Score* de cerca de 99%. O modelo escolhido para ser integrado junto ao sistema obteve um erro de 4.828,8 kg para a RMSE calculada, que pode ser considerado baixo se comparado à proporção dos dados usados.

Palavras-chaves: Apicultura. Aprendizado de Máquina. Regressão. Modelos Preditivos. Mineração de Dados. Apoio à Decisão. Sistema *Web*.

ABSTRACT

It is noticeable the highlight that the beekeeping sector has been gaining in recent years, especially in the production and marketing of honey-related products. In the Brazilian scenario, although we have favorable conditions for the practice of beekeeping, this area suffers from the limited use of technological tools, which directly affects production levels. This paper consists of a 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 (R^2 Score) and the error calculation of the predictions using the Root Mean Squared Error (RMSE). Research results include the development of the proposed Web system and the experimental evaluation of regression methods in different datasets. By using a real scenario dataset, the model using Multiple Linear Regression obtained better performance results when compared to the other models analyzed, obtaining R^2 Score of about 99%. The model chosen to be integrated with the system got an error of 4,828.8 kg for the calculated RMSE, which could be considered low when compared to the proportion of the data used.

Keywords: Beekeeping. Machine Learning. Regression. Predictive Models. Data Mining. Decision Support. Web System.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvores de Decisão
AGs	Algoritmos Genéticos
AM	Aprendizado de Máquina
AR	Análise de Regressão
BI	*Business Inteligence*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CV	*Cross-validation*
HSI	*Health Status Index*
IA	Inteligência Artificial
KDD	*Knowledge Discovery Databases*
LOOCV	*Leave One Out Cross-validation*
MAE	*Mean Absolute Error*
MARS	*Multivariate Adaptive Regression Splines*
MCid	Ministério das Cidades
MD	Mineração de Dados
MDR	Ministério do Desenvolvimento Regional
MEEs	Modelos de Equações Estruturadas
MI	Ministério da Integração Nacional
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
MSE	*Mean Squared Error*
MVC	*Model-View-Controller*
ORM	*Object Relational Mapper*
PCA	*Principal Component Analysis*

PCO	*Predictive models for Colony Outputs*
PoC	*Proof of Concept*
RAE	*Relative Aproximation Error*
RBF	*Radial Basis Function*
RMSE	*Root Mean Squared Error*
RNA	Rede Neural Artificial
RNs	Redes Neurais
SGBD	Sistema Gerenciador de Banco de Dados
SUDAM	Superintendência do Desenvolvimento da Amazônia
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
TCC	Trabalho de Conclusão de Curso
TICs	Tecnologias de Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

Com o grande avanço da ciência e da tecnologia nas últimas décadas, é notório o surgimento de diversos mecanismos e ferramentas criadas pelo ser humano, com a finalidade de fornecer auxílio nas mais variadas tarefas ao longo da vida. Ou seja, a tecnologia está cada vez mais presente em nosso meio e o seu uso tem contribuído de forma positiva para a sociedade, por meio do compartilhamento e disseminação da informação, bem como agilidade, praticidade e eficiência na resolução de problemas. Mais especificamente, as Tecnologias de Informação e Comunicação (TICs) permitem, sobretudo, a informatização de processos e otimização do trabalho, independente da área. Com essa crescente disseminação da informação, um grande volume de dados é gerado e armazenado a todo momento. Dessa forma, sistemas com o intuito de integrar e analisar dados têm surgido. Entretanto, a análise e interpretação dos mesmos devem ser conduzidas de forma adequada e eficaz com o objetivo de auxiliar no processo da tomada de decisão (SILVA; SILVA; GOMES, 2016).

Sistemas de recomendação e sistemas de apoio à decisão, principalmente baseados em Aprendizado de Máquina (AM), têm se mostrado como ferramentas importantes e bastante populares na utilização de dados para auxílio à tomada de decisão em diferentes áreas. Técnicas de Inteligência Artificial (IA), em especial o AM, são constantemente utilizadas com sucesso em diversos problemas reais. Sistemas baseados em AM são capazes de auxiliar na resolução de problemas a partir de inferências sobre um conjunto de dados. Dentre os vários problemas onde a IA tem sido inserida com sucesso nos últimos anos, é possível destacar aplicações de agropecuária, ecologia e meio ambiente, finanças e de saúde (FACELI et al., 2011).

Na área da agropecuária, encontra-se a apicultura, que destina-se à criação e exploração racional de abelhas, praticada pelo pequeno produtor rural ou agricultor familiar, gerando lucros e renda para muitas famílias no mundo (MARANHÃO et al., 2016). A apicultura é uma atividade bastante antiga, onde as civilizações se aproveitavam do uso do mel para fins medicinais (PINTO, 2016). A criação racional de abelhas contribui para a sociedade tanto economicamente quanto ambientalmente. As abelhas são importantes polinizadores e responsáveis por parte dos serviços de polinização que garantem diversos benefícios ao ser humano e ao meio ambiente. Dentre estes benefícios, podemos citar a produção de alimentos, conservação da diversidade biológica e crescimento econômico do país (CERQUEIRA; FIGUEIREDO, 2017).

Portanto, 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). Países asiáticos encabeçam a lista de maiores produtores apícolas apresentando, na última década, um crescimento estável em volume, produção e produtividade por colmeia. A China é a principal concorrente mundial, liderando tanto em produção quanto em exportação de mel (VIDAL, 2017). O Brasil vem se mantendo em uma boa posição no *ranking* mundial já há alguns anos. De 2012 a 2018, o país passou de 33,9 mil para 42,3 mil toneladas de mel produzidas (IBGE, 2018), impulsionado pelo aumento da demanda e pela valorização deste como um produto saudável.

A apicultura no Brasil começou a ganhar expressividade a partir de 1839, quando abelhas de nome científico *Apis Mellifera* foram introduzidas, principalmente nas regiões sul e sudeste do país, por parte de imigrantes europeus (PINTO, 2016). Aliás, ainda hoje, a região sul é a que mais produz mel, concentrando cerca de 43% do total produzido no país. A região nordeste vem logo em seguida, com cerca de 26% de contribuição na produção de mel (IBGE, 2016). As características de clima e flora do nordeste brasileiro permitem à região elevada competitividade no mercado mundial de produtos apícolas. Porém, ainda no Brasil e principalmente no nordeste, o apicultor, pequeno produtor rural, possui baixo nível de profissionalização, dificuldade de acesso à tecnologia e assistência técnica. Além disso, um grande número de apicultores não dispõe de canais de comercialização adequados (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.

O ordenamento apícola é uma ferramenta decisiva para apoio à tomada de decisão. Com ela, os apicultores, além de registrarem as colmeias, o que auxilia no manejo zootécnico, podem administrar seus apiários para potencializar a obtenção do lucro. Ainda, o acompanhamento do sistema produtivo apícola, utilizando ferramentas tecnológicas, como *softwares* de gerenciamento e produção de apiários, além de prover benefícios aos apicultores, pode fomentar políticas públicas a partir das informações atualizadas pelos próprios produtores. Com base nisso, o gerenciamento eficiente da produção apícola a partir de soluções tecnológicas torna-se relevante e tem o intuito de prover a eficiência econômica e a eficácia na execução das atividades

da apicultura.

Neste contexto, existem alguns trabalhos presentes na literatura que buscam auxiliar na gestão e no controle apícola, como em (DUTRA, 2016) e (PINTO, 2016). Contudo, estes trabalhos são, em grande maioria, focados na gestão e no gerenciamento e quase não trazem contribuição no que diz respeito à tomada de decisão. Existem ainda alguns trabalhos que fazem uso de técnicas mais sofisticadas como o AM para auxiliar de alguma forma à produção apícola através da predição de diferentes fatores, como em (GRANDÓN et al., 2016) e (KARADAS; KADIRHANOGULLARI, 2017). No entanto, esses trabalhos levam em consideração as características presentes em outros cenários, o que, por diversos fatores como clima, vegetação, ou os próprios indicativos de produção de um país, diferenciam-se do trabalho desenvolvido no Brasil.

1.1 Motivação

O projeto "Rotas de Integração Nacional"(Portaria MI[1] nº 162, de 24 de abril de 2014) (BRASIL, 2014) estabelece as rotas de integração nacional como estratégia de inclusão produtiva e desenvolvimento regional. Tais rotas constituem "redes de arranjos produtivos locais setorial e territorialmente interligados que promovem a inovação, a diferenciação, a competitividade e a lucratividade dos empreendimentos associados". Exemplos de rotas da integração são as rotas do cordeiro (semiárido nordestino), do açaí (macrorregião norte), do leite (macrorregião centro-oeste e sul), do peixe (macrorregião norte e nordeste), da fruta (macrorregião norte e nordeste) e rota do mel (macrorregião nordeste) (EMBRAPA, 2017a).

Em 2017, representantes de associações de produtores de mel da Amazônia Legal reuniram-se no evento "Oficina de Planejamento da Rota do Mel", em Belém (PA). O evento, organizado pela Superintendência do Desenvolvimento da Amazônia (SUDAM), teve como intuito avaliar 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, 2017c). A oficina contou com a participação de instituições de ensino, pesquisa e extensão, agentes de transferência de tecnologia e o Ministério do Desenvolvimento Regional (MDR) para o planejamento da inclusão da região na Rota do Mel (EMBRAPA, 2017a) (EMBRAPA, 2017b).

Durante a oficina, informações acerca das características da produção do setor foram levantadas a partir da aplicação de questionário, que pode ser conferido no

[1] O Ministério da Integração Nacional (MI) passou a ser incluído dentro do Ministério do Desenvolvimento Regional (MDR), através de sua união com o Ministério das Cidades (MCid), em janeiro de 2019.

Anexo A. Os questionários de rota do mel apresentavam um conjunto de informações referentes a cada associação, cooperativa, federação ou órgão que teve participação. O questionário completo foi composto de 67 questões, divididas em seções de identificação, caracterização e comercialização.

Outro evento que serviu como base para inspiração deste trabalho foi a oficina "Desenvolvimento de plataforma virtual e de aplicativos para apoio à gestão apícola e meliponícola", oficina esta realizada em 2019 através do projeto Agrobio, que visa gerar renda e aumentar a produtividade de atividades agrícolas na região do estado do Pará. A oficina contou com a participação de diversos profissionais do setor, a fim de colaborar com informações importantes para a construção de plataforma eletrônica para organização de cadeia produtiva do mel no estado do Pará.

Nesse contexto, este trabalho surgiu como solução tecnológica para utilização de dados para melhoria da gestão apícola, mediante necessidade demonstrada pela Embrapa Amazônia Oriental[2], com foco no estado do Pará. Esta necessidade pôde ser constatada através dos eventos citados, destacando-se a relevância de uma gestão eficiente da cadeia produtiva. Sendo assim, a proposta deste trabalho abrange o desenvolvimento de uma ferramenta baseada em técnicas de AM para auxiliar na tomada de decisão na gerência dos processos envolvidos no manejo e produção apícola. A partir dos dados adquiridos através desses eventos e *datasets* disponíveis na *Internet* , é possível fazer análises e construir modelos que serão utilizadas para a predição da produção de mel. A ferramenta contará com dois módulos: um módulo *Web* para interação com o usuário e um módulo de inteligência para predições. O sistema será capaz de armazenar os dados relacionados à cadeia de produção, que serão utilizados para predições úteis no processo de tomada de decisão.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo deste trabalho consiste no desenvolvimento de uma ferramenta *Web* inteligente para gestão e predição da produção de mel, utilizando técnicas de AM voltadas para a tarefa de predição a fim de oferecer suporte à tomada de decisão a partir de informações sobre a cadeia apícola cadastradas em uma base de dados.

[2] https://www.embrapa.br/amazonia-oriental

1.2.2 Objetivos Específicos

Para que se alcance o objetivo geral deste trabalho, é destacada abaixo uma série de objetivos específicos:

- desenvolver um sistema *Web* para interação com os apicultores e gestores;
- coletar dados a fim de construir as bases de dados utilizadas neste trabalho;
- analisar e preparar os dados que serão utilizados para o desenvolvimento do módulo de predição, a fim de tornar o processo de aprendizado dos algoritmos mais eficiente;
- estudar e analisar as técnicas de AM que serão utilizadas para a predição da produção de mel;
- mensurar o desempenho de modelos preditivos com base em métricas de avaliação;
- integrar o modelo preditivo com o sistema *Web* para realizar inferências automatizadas sobre a produção.

1.3 Organização do Trabalho

O trabalho está organizado da seguinte forma: O Capítulo 2 trata da fundamentação teórica, onde são abordadas as tecnologias e conceitos utilizados como base para a pesquisa apresentada neste trabalho. O Capítulo 3 aborda os trabalhos relacionados à área de estudo desta pesquisa. O Capítulo 4 descreve a proposta apresentada para este trabalho, bem como a metodologia adotada. O Capítulo 5 explana os resultados obtidos com o desenvolvimento deste trabalho. Por fim, no Capítulo 6, são apresentadas as devidas conclusões, bem como possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

No decorrer deste capítulo, são abordados com mais profundidade, conceitos e tecnologias que compreendem o conteúdo da proposta apresentada, além de servirem como base para o entendimento da mesma.

2.1 Ciência de Dados e Negócios

O desenvolvimento tecnológico em escala mundial tem contribuído cada vez mais para a disseminação de informações. Organizações e empresas precisam, estrategicamente, fazer uso dos dados consumidos e produzidos de forma sábia. É nesse contexto que o uso de sistemas inteligentes que auxiliem em decisões sobre certo domínio se torna um diferencial para quem o possui. O uso de técnicas de análise e exploração podem ajudar a descobrir certas informações e padrões nos dados. Essas informações podem ser úteis para escolher, por exemplo, a técnica de aprendizado a ser utilizada (FACELI et al., 2011). A criação de bases de conhecimento pode ser conquistada a partir de técnicas de Aprendizado de Máquina (AM), Mineração de Dados (MD), e modelos matemáticos, por exemplo (REZENDE, 2003).

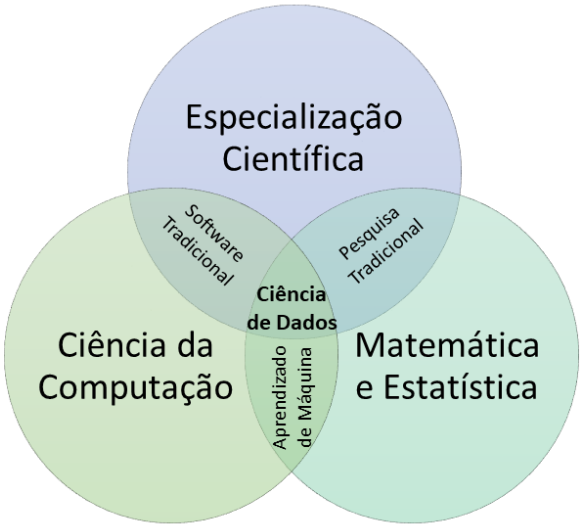
Pequenas e grandes empresas tendem a manipular maiores quantidades de dados com o passar do tempo. Lidar com esse fato pode sobrecarregar os tomadores de decisão de uma organização. Pensando nisso, organizações têm investido cada vez mais em abordagens tecnológicas com o intuito de automatizar essa manipulação. Empresas devem ser capazes de transformar os dados adquiridos em informações relevantes e estratégicas. Logo, a análise e interpretação de grandes quantidades de dados objetiva auxiliar no processo da tomada de decisão em ambientes organizacionais (SILVA; SILVA; GOMES, 2016).

Conceitos e tecnologias de suporte à tomada de decisão têm se popularizado nos últimos 30 anos (LUÍS, 2014). Por exemplo, o *Business Intelligence* (BI) permite o acesso interativo e manipulação dos dados, possibilitando a realização da análise adequada, objetivando otimizar o desempenho da organização (THAMIR; POULIS, 2015). Com o apoio da Ciência de Dados, ou *Data Science*, além de possibilitar a conversão de dados brutos em *insights* de negócios, é possível obter uma análise preditiva e prescritiva dos dados.

Segundo Amaral (AMARAL, 2016), a Ciência de Dados consiste de processos, métodos e tecnologias que tratam de estudar todo o ciclo de vida do dado, desde a sua produção até o seu descarte. De acordo com o Diagrama de Venn, ela se encontra na

interseção de diferentes áreas, como mostrado na Figura 1.

Figura 1 – Ciência de Dados na interseção de diferentes áreas



Fonte: Elaborada pelo autor.

A Ciência de Dados, portanto, tem o intuito de aprimorar a tomada de decisão por meio da análise automatizada de dados. Atualmente, muitas empresas estão investindo em Ciência de Dados utilizando-se de um conjunto de métodos analíticos que compreende a Análise de Regressão (AR). Isso envolve, de forma geral, a previsão e estimação de valores sobre determinados pontos de um negócio. Quando se quer por exemplo, estimar o valor de produção de alguma área do negócio ou encontrar clientes com certo tipo de padrão, são utilizados, dentro do campo da Ciência de Dados, métodos de Aprendizado de Máquina, Estatística Aplicada e Reconhecimento de Padrões (PROVOST; FAWCETT, 2016).

2.2 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é uma subárea da Inteligência Artificial (IA) que busca desenvolver algoritmos capazes de aprender de forma inteligente. Um sistema inteligente de aprendizado baseado em AM é capaz de fazer inferências sobre um domínio de dados e tomar decisões com base na experiência acumulada a partir do sucesso na resolução de problemas anteriores. Este processo é o que caracteriza a indução (FACELI et al., 2011).

A indução, nada mais é do que a inferência lógica sobre um conjunto de exemplos a fim de se obter conclusões genéricas. A inferência indutiva é um dos mais utilizados métodos para se obter novos conhecimentos e predizer algo (REZENDE, 2003).

Um algoritmo de AM busca aprender a partir de um conjunto de treinamento, procurando por uma hipótese indutiva. Cada algoritmo utiliza uma forma de representar as hipóteses. Elas podem ser representadas, por exemplo, em Redes Neurais (RNs), como um conjunto de valores associados aos pesos de cada conexão da rede. Já Árvores de Decisão (AD), por exemplo, utilizam uma estrutura de árvore, onde cada nó interno refere-se ao resultado de uma pergunta correspondente ao valor de um atributo, e cada nó externo está associado a uma classe (FACELI et al., 2011).

Algoritmos de AM podem ser utilizados em diferentes tarefas conforme alguns critérios. Essas tarefas podem ser descritivas, onde os algoritmos seguem o paradigma de aprendizado não supervisionado e buscam encontrar grupos de objetos ou atributos semelhantes no conjunto de dados. Ou ainda, podem ser preditivas, ou seja, para prever um rótulo ou valor para novos exemplos com base nos atributos de entrada, sendo neste utilizada a abordagem de aprendizado supervisionado (FACELI et al., 2011).

Como já citado, a forma de aprendizado dos algoritmos de AM segue alguns paradigmas e são definidos na literatura como mostrado a seguir:

- **Aprendizado não supervisionado**: Neste tipo de aprendizado, os dados utilizados durante o processo de aprendizado dos algoritmos não são constituídos de nenhum tipo de especificação. Os algoritmos precisam aprender padrões e tirar dos dados informações que ajudem a agrupá-los de acordo com suas características (NORVIG; RUSSELL, 2014).

- **Aprendizado supervisionado**: No aprendizado supervisionado, o algoritmo irá aprender com base em um conjunto de atributos de entrada previamente rotulados. O algoritmo irá aprender como esses dados se comportam e então, direcionar um valor de saída a um novo exemplo de entrada (NORVIG; RUSSELL, 2014).

- **Aprendizado semi-supervisionado**: O aprendizado semi-supervisionado é caracterizado pela presença tanto de exemplos de entrada rotulados como também exemplos não rotulados. Além disso, podem haver ainda exemplos rotulados de forma errônea, o que causa ruído nos dados (NORVIG; RUSSELL, 2014).

A partir daqui, são apresentados os modelos de algoritmos preditivos, que são o foco de estudo, em termos de AM, da abordagem proposta neste trabalho.

2.3 Modelos Preditivos

Como descrito anteriormente, dado um conjunto de dados rotulados, um algoritmo de AM preditivo estima valores num domínio conhecido. Se pertencer a um conjunto de valores nominais, tem-se um problema de classificação. Por outro lado, se o domínio pertencer a um conjunto de valores infinito e ordenado, tem-se um problema de regressão. Dados os diferentes tipos de problemas que os modelos preditivos podem atender, será focado então para a abordagem deste trabalho, um problema de regressão.

A tarefa de regressão consiste em modelar a relação entre variáveis numéricas de entrada e saída. As variáveis de entrada são variáveis independentes (não necessariamente independentes entre si), as quais correspondem ao conjunto de atributos utilizados para fazer a estimativa dos valores de saídas. Por sua vez, as variáveis de saída são o conjunto de variáveis dependentes que representam o valor dos atributos que se quer predizer (JOSHI, 2017). Dependendo do problema, diferentes técnicas presentes na literatura podem ser exploradas a fim de encontrar o resultado mais apropriado. A seguir, será detalhado um conjunto de algoritmos de regressão utilizados na realização deste trabalho.

2.3.1 Regressão Linear

Na regressão linear, é assumido que as variáveis se relacionam de forma linear. Sendo assim, com a regressão linear, é possível estimar os dados de um conjunto em forma de uma reta a fim de encontrar o seu melhor grau de inclinação, consequentemente diminuindo o máximo possível os erros entre o valor predito e o valor real. Técnicas de regressão linear podem ser aplicadas dependente da situação. A regressão linear simples, demonstrada pela Equação 2.1, consiste na predição feita a partir de um único atributo preditor, onde y representa a variável dependente, ou seja o atributo alvo da predição e o x o atributo previsor. Já o b_0 e o b_1 representam respectivamente a melhor constante que o algoritmo tentará encontrar, ou seja, de onde a reta irá partir, e o coeficiente para o atributo previsor. Assim, eles indicarão a localização da reta sobre o conjunto de dados. Na regressão linear múltipla é usado mais de um atributo para realizar a predição de um valor. Semelhante a regressão linear simples, na Equação 2.2 mostrada abaixo, tem-se agora um conjunto de variáveis previsoras x_n que terão um coeficiente b_n para cada uma de acordo com o número n de variáveis (PRAKASH; SHARMA; SAHU, 2018).

$$y = b_0 + b_1 * x_1 \tag{2.1}$$

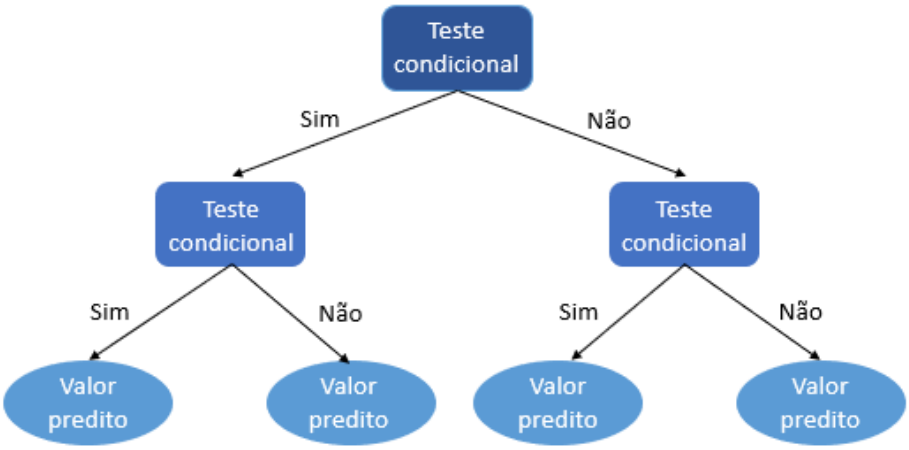
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n \tag{2.2}$$

2.3.2 Decision Tree

Algoritmos de *Decision Tree* partem da premissa de dividir problemas complexos em problemas mais simples e então aplicar uma mesma estratégia para resolvê-los recursivamente em forma de árvore. A árvore é quebrada em conjuntos menores que são definidos através de um teste condicional partindo do nó raiz (topo da árvore) aos nós folha (final de cada ramificação).

Basicamente, essa estrutura de árvore (grafo acíclico) é caracterizada por dois tipos de nós: um nó que deriva um ou mais nós, chamado de nó de divisão; e um nó final, chamado de nó folha. Os nós de divisão contêm testes condicionais referentes aos valores do domínio de um atributo, enquanto os nós folha são rotulados por uma função que leva em conta os valores presentes no atributo alvo. Para os casos de classificação, é usualmente utilizada a função como uma constante que será a moda da amostra dos dados (FACELI et al., 2011). Para problemas de regressão, o algoritmo *Decision Tree* é usado para prever um valor numérico de saída, onde cada nó folha conta com uma função linear de um subconjunto de atributos numéricos, em vez de um único valor (NORVIG; RUSSELL, 2014). A Figura 2 representa a estruturação de uma *Decision Tree* para a tarefa de regressão.

Figura 2 – Estrutura da *Decision Tree* para regressão



Fonte: Elaborada pelo autor.

Para chegar a essa representação, são utilizadas algumas funções. A função de *Entropia*, dada pela Equação 2.3, irá dizer o quanto os dados estão organizados ou desorganizados na base de dados, ou o grau de pureza.

$$Entropia(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2.3}$$

Onde p_i é a probabilidade de cada atributo para uma classe *c*. Essa função é aplicada para cada partição *S* do conjunto de dados. O valor (importância) de um atributo é medido pela função de *Ganho*, dada pela Equação 2.4, que se encarrega de reduzir o valor de *Entropia* encontrado para cada atributo *A*. Assim, será selecionado o melhor atributo a ser utilizado.

$$Ganho(S, A) = Entropia(S) - \sum_{v \in valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.4}$$

2.3.3 Random Forest

O algoritmo *Random Forest* parte da premissa de *Ensemble Learning* (Aprendizado em Conjunto) (MAROS et al., 2019). Esse algoritmo é caracterizado pelo uso de um conjunto de n árvores, onde cada uma resultará em uma decisão e, ao invés de usar a resposta de uma árvore como nos algoritmos simples de *Decision Tree*, o resultado final será a combinação de todas as árvores. Na classificação, diante desse resultado combinado, é escolhido o valor (classe) que mais se repetiu nos resultados das árvores. Em uma tarefa de regressão, em vez disso, usa-se a média dos valores obtidos nas várias árvores utilizadas (WITTEN et al., 2016).

Na implementação desse algoritmo, deve-se escolher o número de árvores que serão utilizadas, sendo ressaltado que a escolha por um número muito alto de árvores durante o treinamento poderá ocasionar num *overfitting*, se ajustando demais àquele conjunto. O *Random Forest* escolhe de forma aleatória um vetor de entrada com k atributos a serem utilizados em cada árvore, garantindo que cada uma irá utilizar atributos diferentes.

2.3.4 Redes Neurais Artificiais

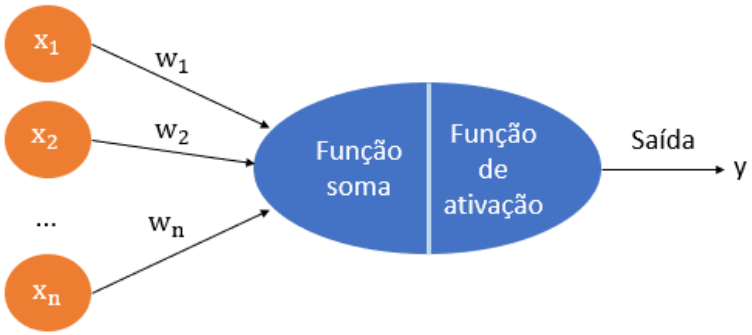
Uma Rede Neural Artificial (RNA) busca simular o comportamento do cérebro humano no processo de aquisição de conhecimento. Similar a um sistema nervoso, sua estrutura é composta por vários neurônios interconectados, formando uma rede. Esses neurônios possuem dendritos, corpo celular e axônio. Analogamente, eles são representados em uma RNA pelas seus valores de entradas, pela função soma e pela função de ativação, respectivamente.

Um dos principais componentes para o aprendizado de uma RNA é o neurônio. Os neurônios de uma RNA são unidades da rede ligados a um ou mais valores de

entrada por meio de terminais que simulam os dendritos de um cérebro. Essas ligações recebem estímulos que exercitam ou inibem aquele valor dependendo do caso, conhecidos como pesos.

RNAs desejam aprender o melhor conjunto de pesos. Os pesos são sinapses que amplificam ou reduzem o sinal de entrada de um neurônio dependendo do seu valor e são atualizados até serem suficientemente pequenos. Os valores recebidos passam por uma função matemática (função soma) que faz o somatório das entradas pelos seus respectivos pesos. A saída de um neurônio é definida por meio de uma função de ativação escolhida dentre várias existentes na literatura de acordo com o problema a ser resolvido (FACELI et al., 2011). Esse valor de saída pode tanto resultar no valor de saída da rede como também pode servir de entrada para outro neurônio de uma nova camada, dependendo se a rede é composta de mais de uma camada oculta. Na Figura 3 é possível visualizar a estrutura de um neurônio.

Figura 3 – Estrutura de um neurônio



Fonte: Elaborada pelo autor.

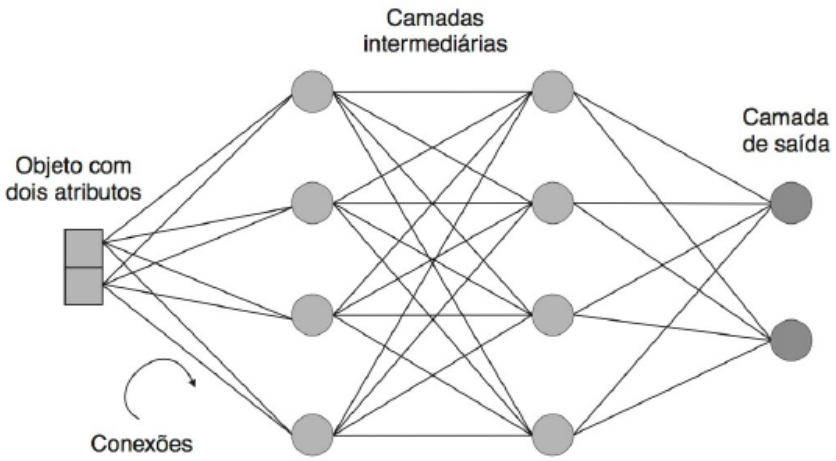
A estrutura de uma RNA é composta de uma camada de entrada, uma ou mais camadas ocultas e um camada de saída. Quando a RNA tem várias camadas ocultas, tem-se uma rede multicamada, onde a saída de um neurônio pode servir como entrada de outros, conforme o total de camadas. Essa rede pode se auto-alimentar, ou seja, valores produzidos na saída de um neurônio pode tanto alimentar a entrada desse próprio neurônio quanto alimentar neurônios de camadas anteriores. Para esse caso, diz-se que existe uma rede de retropropagação ou recorrente. Do contrário, diz-se que uma rede *feed forward* é aquela em que o fluxo de sua alimentação ocorre apenas em um sentido, da camada de entrada à camada de saída (FACELI et al., 2011).

Uma rede *Multilayer Perceptron* (MLP) é uma rede *feed forward* composta de uma ou mais camadas ocultas, como exemplifica a Figura 4. Durante o treinamento usando uma MLP, os pesos são ajustados de acordo com a Equação 2.5.

$$w_i(t+1) = w_i(t) + \alpha\, x_i\, (y_i - \hat{y}_i) \tag{2.5}$$

Onde o peso w_i a ser ajustado no próximo momento t é o resultado do peso no momento anterior somado à multiplicação de uma taxa de aprendizagem, que define a magnitude do ajuste feito, do valor de entrada x_i e o erro calculado $(y_i - \hat{y}_i)$.

Figura 4 – Estrutura de uma rede *Multilayer Perceptron* - MLP



Fonte: Elaborada por (FACELI et al., 2011).

2.3.5 Suport Vector Regression

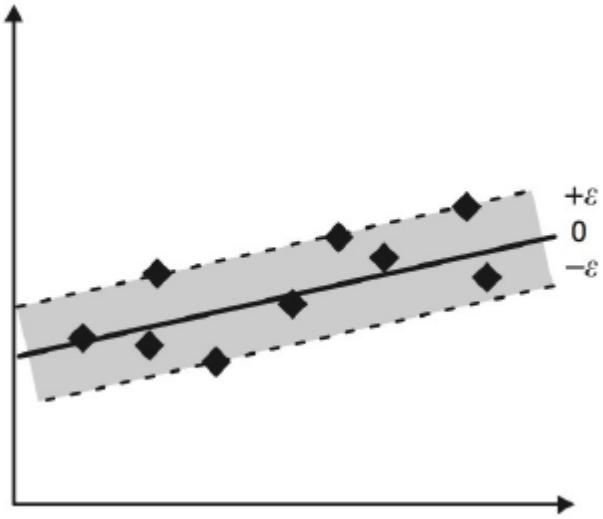
O algoritmo *Support Vector Machine* (SVM) é muito usado pela sua otimização para problemas mais complexos tanto de classificação quanto de regressão. As SVMs são embasadas na teoria do aprendizado estatístico. O algoritmo deve maximizar a probabilidade de uma nova amostra ser predita de forma certa (FACELI et al., 2011). Nessa técnica, o objetivo é aprender hiperplanos com margens máximas. Para tanto, são usados vetores de suportes que são os pontos de cada classe na tarefa de classificação, projetados em um plano, e que o hiperplano que separa os dados seja traçado de forma que a sua distância ao ponto mais próximo de uma classe seja a maior possível. Após serem definidos os vetores de suporte, é desejável que se tenha uma margem elevada, que será usada para a classificação e que cometa poucos erros, minimizando o erro sobre os dados de treinamento, garantindo uma melhor generalização do modelo.

Ao SVM aplicado para tarefas de regressão, damos o nome de *Support Vector Regression* (SVR). Todos os conceitos abordados para a tarefa de classificação também são válidos para a regressão. No entanto, nesse tipo de tarefa, o aprendizado torna-se um pouco mais complexo por se tratar de inúmeros valores numéricos

possíveis na saída. Na regressão, busca-se encontrar um preditor que se ajuste o mais próximo possível dos dados, enquanto que na classificação essa distância não importa, desde que pertença a classe correta. Para medir a qualidade de um preditor, geralmente, utiliza-se uma função perda ou *loss function*. A grande diferença para a tarefa de classificação está no uso de um parâmetro ε, onde a *loss function* definida ignora erros dos valores preditos que estejam dentro de uma distância máxima ε dos valores reais (PRAKASH; SHARMA; SAHU, 2018) com o intuito de minimizar os erros. Assim como na SVM, busca-se encontrar variáveis de folga que consigam lidar com ruídos e *outliers* presentes no conjunto, deixando alguns exemplos de fora da região -ε e +ε. Como ilustrado na Figura 5, a partir da definição dessas variáveis de folga, cria-se uma espécie de tubo onde os erros positivos e negativos são tolerados. A fim de minimizar os possíveis erros que possam existir, tem-se a Equação 2.6, onde ε e $\bar{\varepsilon}$ são as variáveis de folga e C uma constante que impõe o quanto de desvio é tolerado.

$$\frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{n} \varepsilon_i + \bar{\varepsilon}_i) \tag{2.6}$$

Figura 5 – Ilustração do funcionamento da SVR



Fonte: Elaborada por (FACELI et al., 2011).

O emprego mais simples dessa técnica é na aplicação em dados linearmente separáveis. Contudo, pode haver ainda a necessidade de se trabalhar com conjuntos de dados não lineares em que o mapeamento do conjunto de dados passa a ser uma dimensão maior para então poder ser aplicada uma SVM ou SVR linear. Essa dimensão pode às vezes ser muita alta e o processo muito custoso. Para isso faz-se o uso de funções *kernel*. Uma função *kernel* recebe dois pontos como entrada e calcula o produto escalar desses objetos no espaço de características (FACELI et al.,

2011). Algumas das funções mais utilizadas são a linear, a polinomial e a *Radial Basis Function* (RBF), avaliadas neste trabalho.

2.4 Análise de Componentes Principais

Quando se trabalha com dados que representam muitas dimensões, muitas vezes é necessário reduzir essa dimensionalidade para encontrar, de fato, a variação dos dados. Uma das técnicas mais utilizadas é a Análise de Componentes Principais (*Principal Component Analysis* - PCA). Essa técnica consiste em redimensionar o tamanho do vetor de entrada, composto de um conjunto de variáveis inter-correlacionadas, correlacionando os exemplos estatisticamente (ABDI; WILLIAMS, 2010). Com o uso dessa técnica é possível eliminar as características menos importantes e capturar a parcela do conjunto que melhor represente a variação dos dados (FACELI et al., 2011; GRUS, 2018).

2.5 *GridSearchCV*

Esta técnica, disponibilizada pela biblioteca *scikit-learn* (PEDREGOSA et al., 2011), permite a avaliação de modelos, sobretudo a escolha dos melhores parâmetros a serem utilizados pelos modelos criados a partir dos algoritmos de AM, utilizando o conceito de *Cross-validation* (CV). Dentro do *GridSearch*, durante o treinamento do estimador, são selecionados os parâmetros através de uma *grid* de parâmetros definidos pelo especialista, aprimorando o desempenho do modelo (PEDREGOSA et al., 2011). O modelo generalizado poderá então ser usado posteriormente por qualquer conjunto de dados.

2.6 Avaliação dos Modelos Preditivos

A validação dos modelos preditivos gerados a partir do uso das técnicas de AM envolve, geralmente, a execução de experimentos dentro de um ambiente controlado. Nesses experimentos são usados diversos procedimentos que buscam avaliar o desempenho do algoritmo, validando seus resultados. É o caso das métricas de erro e técnicas de amostragem.

2.6.1 Cross-validation

Técnicas de amostragem consistem em dividir o conjunto de dados em subconjuntos de treinamento e de teste a fim de obter estimativas mais confiáveis. A *K-fold*

Cross-validation é uma técnica de amostragem que consiste em dividir o conjunto de dados em K partições de tamanhos aproximadamente iguais. Em cada iteração da CV, uma partição é deixada de fora do conjunto de treinamento e utilizada no conjunto de teste. Esse processo é repetido K vezes, utilizando a cada iteração uma nova partição formada aleatoriamente. O resultado final é gerado ao contabilizar a média do desempenho de todos os conjuntos de teste a partir das métricas de avaliação. É observado ainda que um desvio padrão muito alto indica uma alta instabilidade do modelo, podendo este ser usado como critério de desempate para modelos com desempenhos muito parecidos (FACELI et al., 2011).

Outra técnica utilizada, dependendo do problema, é a *Leave One Out Cross-validation* (LOOCV) que, por ter um custo computacional muito alto, é mais apropriada para bases menores, produzindo estimativas mais confiáveis. Ela segue o mesmo princípio da CV convencional, no entanto, o número de partições geradas é equivalente ao total de amostras. Em cada iteração da LOOCV, um único exemplo é deixado de fora do treinamento e utilizado no conjunto de teste (FACELI et al., 2011).

2.6.2 Coeficiente de Determinação

O coeficiente de determinação, também conhecido como R^2 ou *Score*, é uma métrica que permite medir o quanto o modelo avaliado está se ajustando aos dados, ou seja, o quanto de variabilidade dos valores preditos (y) é explicado em função da variação dos atributos previsores (x). Essa métrica mede a relação entre a variação total e o quanto dessa variação não é descrita pela reta projetada. Ela pode ser demonstrada pela Equação 2.9, onde S_{TOT}, descrita pela Equação 2.7, é o somatório do quadrado dos erros em relação à média ($\bar{y}$) dos valores de y e S_{RES}, descrita pela Equação 2.8, é o somatório dos erros ao quadrado (MILES, 2014).

$$S_{TOT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2.7}$$

$$S_{RES} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.8}$$

$$R^2 = 1 - \frac{S_{RES}}{S_{TOT}} \tag{2.9}$$

O seu resultado pode variar, geralmente, entre 0 e 1, onde 0 significa que o modelo teve desempenho semelhante à reta da média dos valores de y, e 1 significa que o modelo se ajustou perfeitamente aos dados. É importante salientar ainda que um resultado negativo dessa métrica indica um péssimo desempenho do modelo pre-

ditivo, o que pode ser interpretado como um resultado pior que a média dos dados de y.

2.6.3 Cálculo do Erro

As métricas de cálculo do erro, de maneira simples, calculam o quanto de ruído foi gerado a cada resultado da predição realizada pelo modelo. Para casos de regressão, é possível verificar o erro de uma predição ao calcular a distância entre um valor predito $\hat{y}$ e o valor real y. Métricas bastante abordadas na literatura são a *Mean Absolute Error* (MAE), a *Mean Squared Error* (MSE) e a *Root Mean Squared Error* (RMSE).

A métrica MAE, apresentada na Equação 2.10, calcula apenas o valor absoluto dado pelo somatório dos erros pela quantidade de exemplos n.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10}$$

A segunda métrica abordada é a MSE, que diferente da MAE, penaliza os erros maiores através da Equação 2.11, onde o calculo é realizado através do somatório dos erros ao quadrado dividido pela quantidade de exemplos n.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.11}$$

A RMSE é uma adaptação da métrica MSE, pois além de penalizar os erros maiores e dar menos importância a erros menores, o que aumenta consideravelmente a proporção do erro, é calculada a raiz quadrada para que o valor obtido volte à mesma proporção dos dados utilizados. Essa métrica é descrita pela Equação 2.12.

Por esse motivo, a RMSE foi utilizada para avaliar os modelos, uma vez que ela permite uma fácil interpretação e o erro volta a estar na mesma unidade de medida dos dados. Por exemplo, um modelo que deveria prever 1000 kg, previu 990 kg. A RMSE permite verificar que o erro foi de 10 kg.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.12}$$

3 TRABALHOS RELACIONADOS

Os trabalhos apresentados neste capítulo correspondem a abordagens presentes na literatura que buscam auxiliar o setor apícola de alguma forma, seja por meio de auxílio na questão gerencial ou com métodos que buscam ajudar na produtividade e saúde das colmeias.

No trabalho de (GILIOLI et al., 2018) são propostas duas modelagens de abordagem estatística baseadas em Modelos de Equações Estruturadas (MEEs) para avaliação da produtividade e saúde das colônias de abelhas. A primeira delas apresenta um modelo *Health Status Index* (HSI) para estimar uma relação de causa-efeito entre variáveis latentes. É definido um modelo conceitual composto por uma síntese de conjuntos de variáveis latentes dispostos de forma gráfica por dois submodelos: um estrutural, definindo as relações entre as variáveis, e outro de medição, definindo as ligações entre as variáveis e seus respectivos indicadores.

Ainda sobre o trabalho de (GILIOLI et al., 2018), a segunda modelagem, *Predictive models for Colony Outputs* (PCO) permite, dado um conjunto de dados de entrada, predizer a produtividade e os serviços de polinização fornecidos pelas abelhas com base no modelo HSI. Foi utilizado o algoritmo Regressão Linear sobre uma base de dados dividida em 75% para treinamento e 25% para teste a fim de verificar a relevância de cada variável sobre o valor estimado de produção das abelhas, bem como a capacidade de satisfação dos serviços de polinização. Os testes preliminares se deram a partir de um *dataset* representando condições reais as quais os apicultores encontram em diferentes áreas da Grécia. Com base nisso, a fase de análise contou com a participação de especialistas da apicultura para expressarem estimativas sobre produção e distribuição de atributos indicadores da colônia (Rainha, Contaminação, Comportamento e fisiologia, entre outros). Partindo dessas incertezas, foi gerada, aleatoriamente, uma base de dados com 1000 observações. Para essas análises, foi utilizada a linguagem R juntamente com o pacote plspm. A partir dos resultados das análises e experimentos sobre a base de dados, foi possível perceber o nível de significância entre os atributos da colônia e fatores que têm influência sobre eles. Entre estes resultados foi possível comprovar uma forte relação entre a saúde da colônia de abelhas e a quantidade de mel colhida.

As principais diferenças entre a abordagem apresentada em (GILIOLI et al., 2018) e a deste trabalho podem ser enumeradas a seguir: (i) diferentes usos de modelos de predição. Enquanto este trabalho foca em uma abordagem usando métodos de AM, (GILIOLI et al., 2018) fazem uso de modelos estatísticos para fazer as análises; (ii) o foco das análises baseia-se em características presentes na Europa; (iii) os

autores não abordam a implementação de um módulo que interaja com um usuário de modo a fornecer funcionalidades de gerenciamento apícola.

Na proposta de (PINTO, 2016), são levantadas informações sobre a gestão dos apiários, aplicando-se um questionário a apicultores de diferentes associações do estado de São Paulo. Os resultados desse questionário, obtidos em forma de porcentagem, indicam a situação geral do apicultor no que diz respeito ao conhecimento do empreendimento e uso da informação na gestão do apiário. Com isso, é possível gerar indicadores necessários para a tomada de decisão. Foi possível constatar, na análise dos resultados, que os apicultores mostravam carência no registro das informações. Para esse problema, fez-se o uso de planilhas eletrônicas nas quais os apicultores guardam informações e é retornado, por exemplo, o cálculo de previsão da produção projetada, além de outros fatores. A partir disso, foi construído um aplicativo com a ferramenta Fábrica de Aplicativos, utilizado para coleta das informações no campo e as mantendo, minimizando as perdas de informações. Também foi desenvolvido um *Website* usando Wix para auxílio informativo e organização dos processos pertencentes à gestão apícola.

A pesquisa discutida em (PINTO, 2016) não apresenta inovações no que se refere à aplicação de técnicas para inferências a partir dos dados coletados, como a aplicação de métodos de AM, proposta neste trabalho. Foi observada também a opção pelo uso de ferramentas simplistas e limitadas para desenvolvimento das aplicações propostas, o que implica na não exploração de conhecimentos mais aprofundados de programação.

Em (GRANDÓN et al., 2016), é apresentada a proposta de uma ferramenta para predição de florada e tomada de decisão para a produção de mel orgânico a fim de melhorar os planejamentos de manejo produtivo da colmeia. O desenvolvimento foi focado na região de *Maule*, no Chile. A ferramenta é composta por um sistema de informação que conta com três módulos: 1) gestão de apiários; 2) cartografia digital de zona geográfica de interesse; e 3) predição de florada, baseada em técnicas de IA e dados climatológicos. O módulo de cartografia mantém um mapa com dados a respeito de pontos geográficos de apiários, localização dos cultivos transgênicos da região e estações meteorológicas.

Em se tratando do módulo de predição, os modelos desenvolvidos por (GRANDÓN et al., 2016) foram baseados em RNAs. Uma MLP foi utilizada para o treinamento de cada modelo. O módulo de predição de florada tem o objetivo de inferir a data aproximada, em dias, do começo de uma florada para uma espécie botânica de interesse apícola, com base em dados históricos das espécies associadas à estação meteorológica mais próxima. Os melhores resultados para o treinamento foram obtidos com a configuração de 3 camadas compostas de 6 neurônios cada. A camada de en-

trada contou com um conjunto de valores correspondentes à temperatura, radiação e umidade. Esses valores foram otimizados durante o treinamento através do uso de Algoritmos Genéticos (AGs) que se encarregaram de fazer os ajustes dos pesos. Para cada uma das espécies botânicas, foram considerados diferentes modelos preditivos. Tais modelos foram avaliados utilizando a técnica LOOCV. A partir dos resultados da LOOCV, a função de erro utilizada foi a RMSE. Foram obtidos, portanto, erros preliminares próximos a 7 dias, que puderam ser considerados baixos dentro do período de medições de uma semana.

A partir do exposto acima, o trabalho de (GRANDÓN et al., 2016) apresenta algumas diferenças em comparação com a proposta apresentada neste trabalho: (i) os locais de escolha para estudo e desenvolvimento das propostas apresentam características diferentes e os resultados obtidos com as abordagens podem não satisfazer a realidade de cada ambiente; (ii) aliás, os resultados apresentados ainda não são definitivos; (iii) apesar da abordagem de (GRANDÓN et al., 2016) utilizar métodos semelhantes para predição baseados em AM, os autores focam suas predições em volta dos dias de início de florada, enquanto a abordagem deste trabalho foca na previsão do nível de produção de mel.

No trabalho desenvolvido por (KARADAS; KADIRHANOGULLARI, 2017), foi realizado um estudo censitário para coletar informações de 85 (tamanho da amostra) fazendas de apicultura em Igdir, na Turquia, em 2014. O objetivo dessa pesquisa esteve em identificar fatores que influenciam a produção de mel por colmeia. Para esse propósito, foram levados em considerações a idade dos apicultores (A_1), nível de educação (A_2), número de colmeias cheias (A_3), espécie das abelhas (A_4), o tempo gasto no planalto (A_5), alimentação no outono e primavera (A_6), período de trabalho no ano (A_7), frequência de troca da rainha (A_8) e controle da colmeia no verão (A_9). Nesse contexto, avaliou-se a performance comparativa dos algoritmos CART, CHAID, *Exhaustive* CHAID, MARS (*Multivariate Adaptive Regression Splines*) e a Rede Neural Artificial *Multilayer Perceptron* (MLP) a fim de encontrar dentre os algoritmos, o de melhor acurácia preditiva com relação ao rendimento médio de mel. A comparação estatística do desempenho dos algoritmos foi executada seguindo as métricas Coeficiente de Determinação (R^2), Coeficiente de Determinação Ajustado ($Adj.R^2$), Coeficiente de Variação, Desvio Padrão (SD_{ratio}), *Relative Approximation Error* (RAE), e *Root Mean Squared Error* (RMSE).

As configurações dos experimentos propostos por (KARADAS; KADIRHANOGULLARI, 2017) se deram utilizando *Cross Validation* com 10-*folds* alocados randomicamente em aproximadamente 8 partes iguais para os algoritmos CART, CHAID, *Exhaustive* CHAID e MARS. Já para o algoritmo MLP, a base de dados foi dividida em 80% para treinamento e 20% teste, usando o método *Holdout*, com 1 camada

oculta composta de 3 neurônios. A função de ativação usada na camada de saída foi a tangente hiperbólica. Foram consideradas significantes, portanto, apenas as variáveis (A_3), (A_7) e (A_9) no algoritmo *Exhaustive* CHAID, as variáveis (A_1), (A_3) e (A_7) no algoritmo CART e no algoritmo MARS foram consideradas apenas as variáveis (A_1), (A_2), (A_3), (A_5), (A_6), (A_7) e (A_8). Sobre as análises e experimentos realizados, ficou concluído então, que o algoritmo MARS teve melhor desempenho quando comparado com os demais algoritmos.

Diante do exposto acima, as diferenças encontradas entre o trabalho de (KARADAS; KADIRHANOGULLARI, 2017) e este trabalho estão em: (i) o trabalho se concentra na realização dos experimentos de AM e não conta com o desenvolvimento de uma plataforma para interação com o usuário. Partindo disso, (ii) esse trabalho comparado não relata nenhum mecanismo de integração com um serviço *web* por exemplo, para o modelo preditivo encontrado.

Diante dos trabalhos apresentados, como pôde-se perceber, a literatura é riquíssima de soluções tecnológicas que buscam auxiliar a eficiência dos processos envolvidos na gerência adequada da apicultura de modo a otimizar a produção e seus aspectos gerenciais. A realização deste trabalho busca contribuir cientificamente com o uso de abordagem inteligente, fazendo uso de AM focando no aproveitamento dos dados para realização de predições e tomada de decisão no contexto apícola. Este trabalho tem o seu foco no cenário brasileiro, utilizando dados reais do estado do Pará. Portanto, os modelos gerados aqui são fiéis ao contexto trabalhado, revelando características distintas para aplicação, especificamente, no Norte e Nordeste.

4 Proposta

O desenvolvimento dessa proposta consiste de uma ferramenta Web para gestão apícola baseada em técnicas de AM para predição de produção de mel, uma vez que foi demonstrada a carência de mecanismos tecnológicos que auxiliem no processo de gerência e tomada de decisão no que diz respeito à produção de produtos apícolas no cenário brasileiro.

A proposta de um sistema Web para gestão apícola torna-se relevante por fornecer uma solução capaz de ser utilizada em qualquer lugar, bastando utilizar um dispositivo que possua conexão com a *Internet*, como computadores, *notebooks*, *tablets* e *smartphones*. A ferramenta fornece ao produtor rural um *software* que facilitará a tomada de decisões na gestão de produção de mel. A adição da funcionalidade de inferências baseadas em AM torna diferenciada a aplicação proposta no momento em que os dados relacionados à cadeia de produção apícola, armazenados no sistema, são utilizados para construção de modelos preditivos. A interface *Web* conta ainda com recursos de geolocalização para visualização de propriedades cadastradas de forma a constituir uma informação relevante a ser usada por gestores e produtores.

Esta pesquisa possui dois fluxos de trabalhos. O primeiro envolve as etapas mostradas na seção anterior com relação à preparação e análise de dados, modelagem, e avaliação de modelos preditivos, utilizando aprendizado supervisionado em uma tarefa de regressão com o objetivo de predizer a produção de mel. O segundo diz respeito ao desenvolvimento de um sistema *Web* (*frontend* e *backend*) como prova de conceito (*Proof of Concept* - PoC). Este conta com recursos para a gestão apícola e geolocalização, como forma de armazenar informações importantes para a criação de inferências baseadas em AM através do modelo avaliado, consumido pelo sistema.

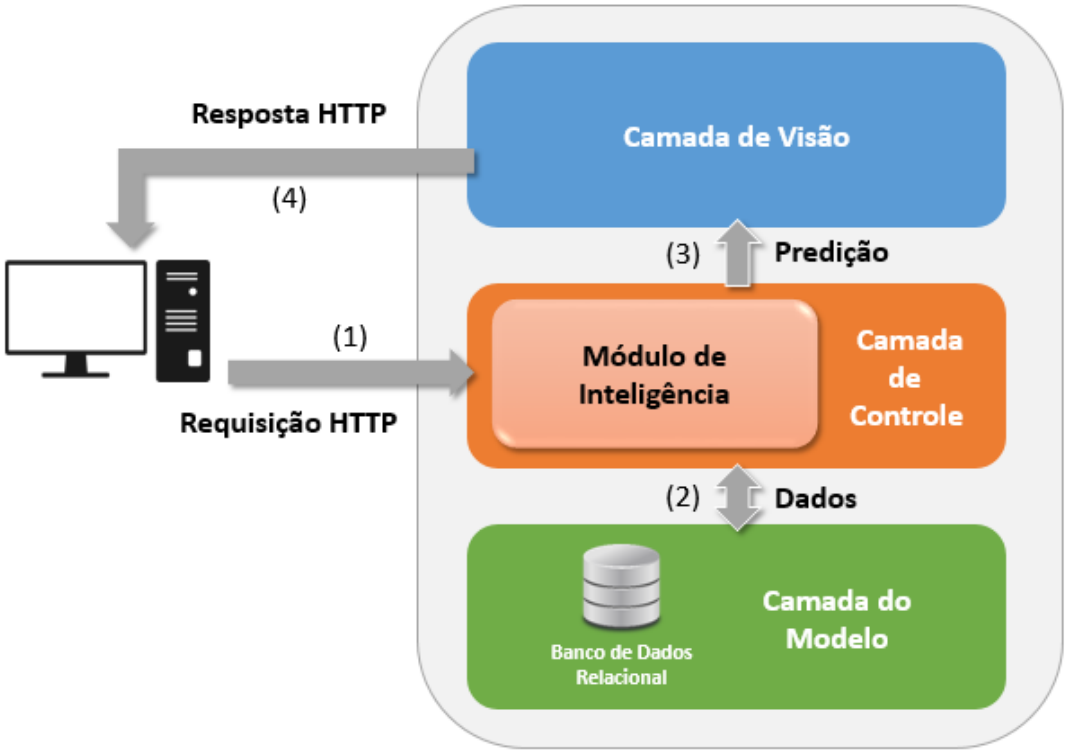
A arquitetura da solução proposta baseia-se no padrão de arquitetura de *software Model-View-Controller* (MVC). Esse padrão consiste em dividir o desenvolvimento de um sistema em três camadas: camada de visão, camada de controle e camada do modelo (SOMMERVILLE, 2011).

A camada de visão é a responsável por fazer a interação com o usuário. É nela que são apresentadas as páginas por meio de arquivos html. A camada de controle recebe as requisições feitas e aciona métodos específicos, podendo se comunicar com a camada do modelo, buscando os dados quando necessário, e com a camada de visão, para onde são repassados os dados requisitados para exibição ao usuário.

A arquitetura da ferramenta proposta é ilustrada pela Figura 6, onde em (1) é feita uma requisição HTTP à camada de controle, em busca de realizar uma pre-

dição 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 6 – Arquitetura da solução proposta



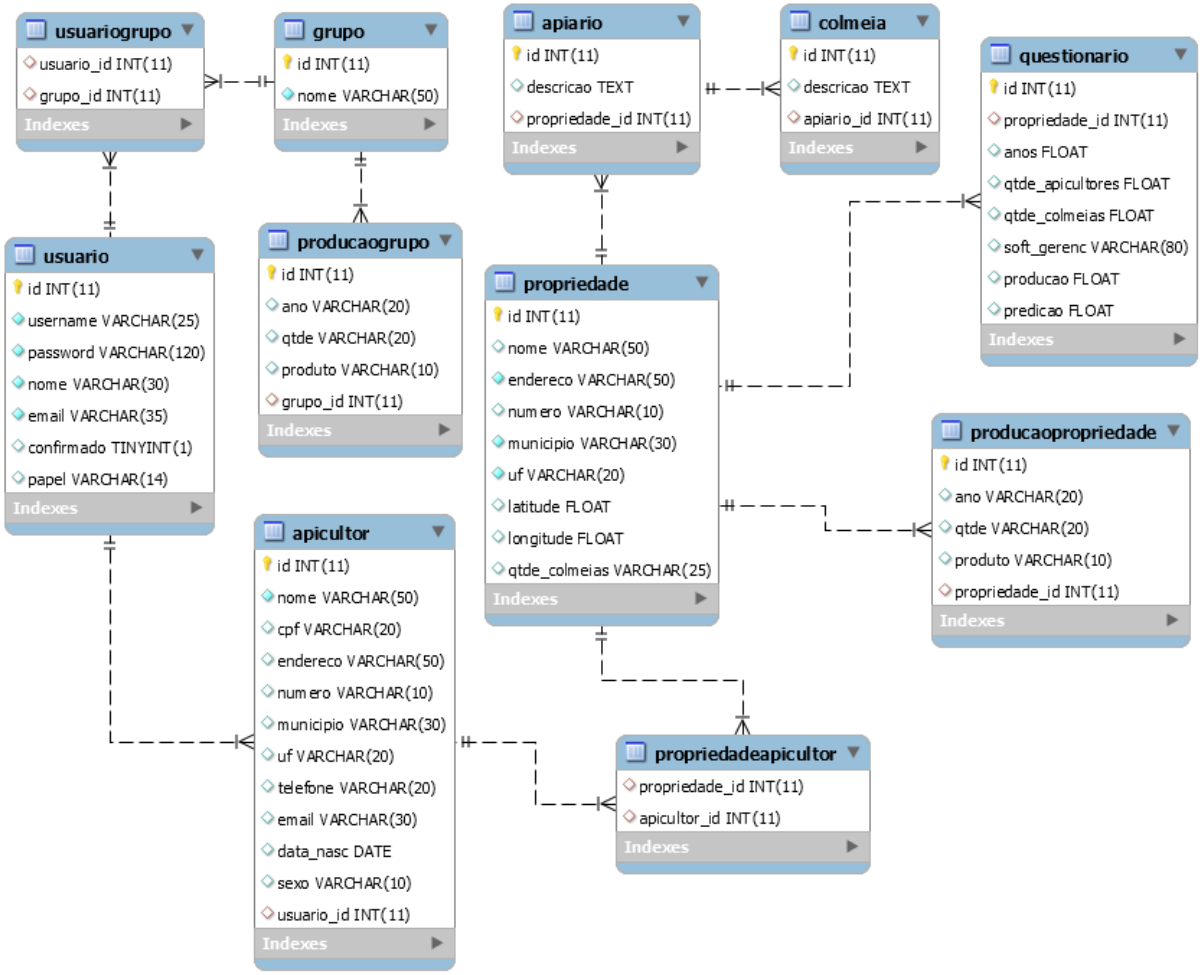
Fonte: Elaborada pelo autor.

Optou-se por utilizar neste trabalho, tanto no sistema *Web* quanto no módulo de inteligência, a linguagem de programação Python 3[1] como linguagem principal de desenvolvimento, com o auxílio do *microframework* Flask[2] para o desenvolvimento *Web*. A linguagem Python já se mostrou como uma das linguagem mais populares entre os cientistas de dados por, além de ser simples e de fácil codificação, fornecer diversas bibliotecas úteis no processo de aprendizado (GRUS, 2018). O acesso aos dados se deu por meio da conexão do Flask com o Sistema Gerenciador de Banco de Dados (SGBD) MySQL[3] através do *Object Relational Mapper* (ORM) SqlAlchemy[4], como forma de mapear os objetos em linguagem Python. A ferramenta consiste de um conjunto de entidades que compõem o banco de dados relacional da aplicação,

[1] http://python.org/
[2] http://flask.pocoo.org/
[3] https://www.mysql.com/
[4] https://www.sqlalchemy.org/

armazenando os dados referentes ao cenário em questão. A Figura 7 apresenta a modelagem do banco de dados utilizado na ferramenta.

Figura 7 – Modelagem do banco de dados



Fonte: Elaborada pelo autor.

O usuário cadastrado no sistema *Web* para gestão apícola poderá acessá-lo por meio de login e senha. A ferramenta permitirá a utilização por parte de gestores e apicultores presidentes de associações e/ou cooperativas, além de pequenos produtores donos de propriedades rurais. Ao acessar o sistema, o usuário será capaz de manter o cadastro de suas propriedades bem como gerenciar seus apiários, colmeias e uma rede de apicultores vinculados à sua propriedade.

Após a inserção das informações das propriedades pelo usuário, serão utilizadas as coordenadas capturadas pelo endereço informado para que sejam visualizadas em forma de marcadores dispostos em um mapa. A funcionalidade de geolocalização tem a capacidade de auxiliar os apicultores na percepção do cenário apícola através da visualização da distribuição geográfica das propriedades presentes em sua

região. Ela foi desenvolvida a partir do uso da biblioteca Folium[5], do Python. Essa biblioteca possibilita a geração automática de mapas a partir dos dados passados em Python, sendo possível visualizá-los com o auxílio da biblioteca Leaflet[6]. A Folium se encarrega de gerar o arquivo html contendo o conteúdo do mapa para a visualização. Com a Folium, foi possível passar as coordenadas iniciais (par contendo a latitude e longitude) indicando onde o mapa será projetado e, através das coordenadas das propriedades informadas, inserir marcadores que representam suas localizações.

Ademais, a ferramenta permite o preenchimento de formulário com informações a respeito das propriedades rurais e sua produção de mel. Essas informações alimentam o banco de dados do sistema e são de fundamental importância para a criação de uma base de dados histórica, a qual contribuirá para a realização da predição a partir de um modelo de AM treinado. Através de algoritmos de AM voltados para regressão, o modelo de predição será capaz de predizer a produção do mel para, com base nisso, servir como auxílio para a gestão de seus níveis de produção e tomada de decisão.

4.1 Metodologia

A partir do que já foi exposto até o momento, esta seção descreve a metodologia utilizada para a realização dos objetivos apresentados no início deste trabalho, descrevendo principalmente os procedimentos envolvidos na construção dos modelos preditivos e o desenvolvimento da ferramenta *Web*.

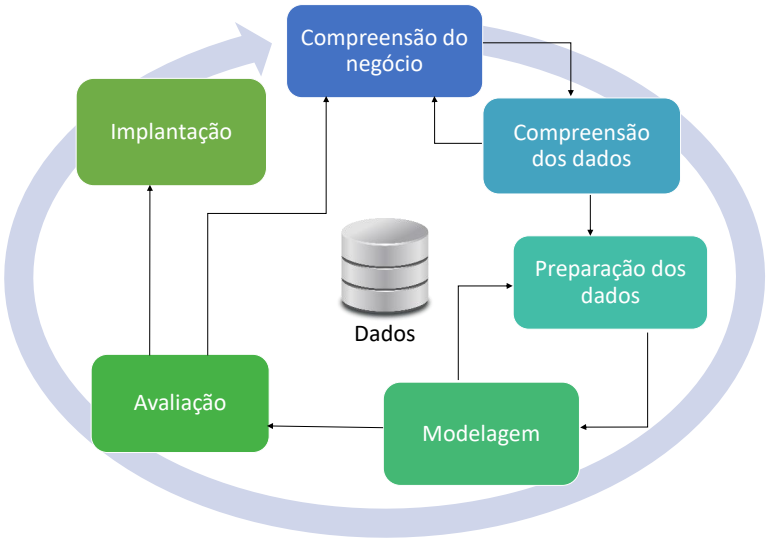
4.2 Etapas de Mineração de Dados

A Mineração de Dados (MD) fornece um conjunto de técnicas que busca explorar e analisar os dados a procura de padrões e correlações, produzindo *insights* e auxiliando na tomada de decisão (WITTEN et al., 2016). Com a popularização do processo de minerar dados, surgiram alguns modelos que são seguidos em projetos que usam MD, como o *Knowledge Discovery Databases* (KDD) e o *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Estes modelos auxiliam no processo de descoberta de conhecimento seguindo um conjunto de passos. Neste trabalho, decidiu-se seguir a estrutura adotada pelo CRISP-DM, indo desde o entendimento do negócio, posteriormente o entendimento dos dados a se trabalhar, seguido das fases de preparação dos dados, modelagem e avaliação. Essa estrutura segue um ciclo onde é possível voltar à fases anteriores se assim for necessário, como pode ser visto

[5] https://python-visualization.github.io/folium/
[6] https://leafletjs.com/

na Figura 8. Dependendo do resultado analisado na fase de avaliação, é possível que se tenha que voltar a compreensão do negócio e rever o que pode ter sido despercebido. Ou ainda, se o resultado for satisfatório, encerrar o ciclo da MD, partindo para o desenvolvimento (PROVOST; FAWCETT, 2016). A seguir, são apresentadas algumas etapas presentes neste trabalho, baseadas nesses modelos de processo.

Figura 8 – Etapas de Mineração de Dados do CRISP-DM



Fonte: Elaborada pelo autor.

No início da tarefa de construção dos modelos preditivos, foram realizadas etapas da mineração de dados com dados disponibilizados no repositório *online* Kaggle. O objetivo se deu em verificar a possibilidade de encontrar um modelo preditivo satisfatório para o problema. Posteriormente, foram realizados todas as etapas apresentadas neste trabalho para um *dataset* mais fiel ao cenário em questão.

4.2.1 Aquisição dos Dados

Os primeiros experimentos correspondem aos dados de um *dataset* disponível publicamente no repositório *online* Kaggle[7]. Inicialmente, foi selecionado o *dataset Honey Production in the USA*, que aborda dados sobre a produção de mel nos Estados Unidos entre os anos de 1998 e 2012. Esse *dataset* foi impulsionado pelo declínio da população de abelhas, consequentemente diminuindo o nível de produção interno do país. O *dataset* é composto de 8 atributos descritos na Tabela 1, sendo o atributo em negrito o alvo da predição, e um total de 626 amostras.

Além da utilização de dados públicos disponíveis na *internet*, também foram utilizados dados disponibilizados pela Embrapa Amazônia Oriental. Esses dados fo-

[7] https://www.kaggle.com/jessicali9530/honey-production

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 (resultado do produto do numcol x yieldpercol), dado em libras.
stocks	Ações detidas pelos produtores, dadas em libras.
priceperlb	Preço médio por libra com base nas vendas expandidas, dado em dólares.
prodvalue	Valor de produção (resultado do produto totalprod x priceperlb), dado em dólares.
year	Ano o qual o dado pertence.

ram coletados através da aplicação de questionários no evento "Oficina de Planejamento da Rota do Mel", ocorrido em 2017.

O questionário, disponível no Anexo A, conta com 67 questões, divididas por meio de seções que dizem respeito à identificação e caracterização de órgãos, associações, federações ou cooperativas, e a comercialização feitas por elas. A partir das perguntas e respostas informadas nos questionários foi possível criar um *dataset*, onde cada pergunta transformou-se em um atributo. O *dataset* completo é composto de 231 atributos e um total de 23 exemplos. Posteriormente, será detalhada a construção de diferentes *datasets*, derivados deste, para a realização de diferentes experimentos.

Na Tabela 2 podem ser conferidas as características originais de cada *dataset*.

Tabela 2 – Características originais dos *datasets* utilizados

Dataset	**Nº de exemplos**	**Nº de atributos**	**Média (atributo alvo)**	**Desv. Padrão (atributo alvo)**	**Min. (atributo alvo)**	**Max. (atributo alvo)**
Kaggle	626	8	4.169.086,26 lb	6.883.846,75 lb	84.000 lb	46.410.000 lb
Embrapa	23	231	34.233,33 kg	67.783,43 kg	15 kg	250.000 kg

4.2.2 Preparação dos Dados

Antes que os dados coletados sejam analisados e explorados, eles precisam passar por uma fase de preparação ou pré-processamento. Técnicas desse tipo aju-

dam a tratar problemas de inconsistência, dimensionalidade, ou problemas com dados incorretos e faltantes, o que melhora a qualidade dos dados (FACELI et al., 2011).

O *dataset Honey Production in the USA* possui boa qualidade com relação aos dados disponibilizados. Por essa razão, para esse *dataset*, não foi necessário tanto empenho na fase de preparação além da transformação do atributo *state*, que originalmente era um tipo de atributo categórico, o qual descrevia a sigla do estado. A fim de obter um melhor desempenho dos algoritmos de AM, foi necessário transformar este em atributo numérico. O método utilizado para isso foi o *LabelEncoder* da biblioteca *scikit-learn*. Esse método consiste em transformar cada valor diferente presente no atributo em um valor numérico correspondente ao valor original. Assim, o tipo do atributo passa a ser numérico.

Para o segundo cenário, primeiro foi construído um *dataset* a partir dos questionários obedecendo a originalidade das respostas. Foi necessário utilizar uma estratégia para questões de múltiplas escolhas em que era possível marcar mais de uma opção a fim de cada resposta consistir apenas em um valor. A estratégia adotada simula o que em AM é chamada de variável *dummy*, onde cada valor possível para a pergunta se torna um atributo, atribuindo 1 ou 0 para o caso de estar marcado ou não, respectivamente. A partir disso, o *dataset* criado conta com 231 atributos e apenas 23 amostras.

Após a criação da primeira versão do *dataset*, se fez necessário criar um dicionário de dados contendo todas as informações que descrevem cada atributo e seus possíveis valores. Isso facilitou a identificação do que cada atributo representa.

Uma das maiores dificuldades encontradas foi a limitação apresentada pelos dados adquiridos. Além de serem poucos, alguns questionários apresentavam uma má qualidade nas respostas fornecidas como por exemplo, uma grande quantidade de perguntas deixadas sem resposta. Além disso, logo inicialmente, pôde-se perceber a irrelevância de alguns atributos para a realização dos experimentos. Por todos esses motivos, foi realizada a remoção dos atributos irrelevantes bem como atributos com poucos ou nenhum dado.

Alguns atributos também apresentavam valores inconsistentes, como no caso dos atributos C54 (Custo de aquisição (R$/kg) de compra do mel do apicultor) e C57 (Preço de venda (R$/kg) do mel fracionado (beneficiado) para associações e cooperativas). No caso do atributo C54, foi encontrado um valor correspondente ao kg em vez do valor em R$. Nesse caso, esse valor foi substituído por um valor nulo. Para o atributo C57, em vez do valor específico em R$, foi detectada uma faixa de preço que correspondia ao valor "90-100", sendo substituído pela média desse intervalo. Também foi feito o tratamento onde valores do tipo *float* estavam como *string*. Além disso, assim como no *dataset* do cenário anterior, realizou-se a transformação de atri-

butos categóricos em atributos numéricos. Com isso, foi gerada uma nova versão do *dataset*, consistindo agora de 206 atributos.

Com essa nova versão, foi verificada a existência de valores nulos e feito o tratamento adequado. Nesse tipo de tratamento, através do método *Imputer* do *scikit-learn*, foi possível definir a estratégia utilizada na substituição dos valores faltantes. A estratégia escolhida foi a média para atributos com valores contínuos. Para atributos de valores discretos, a estratégia utilizada foi adicionar o valor "Não respondeu" às perguntas deixadas em branco.

Após o tratamento adequado, os dados foram submetidos à etapa de processamento. Nesta fase, foram aplicados diferentes algoritmos de AM para realizar o treinamento sobre os dados, o que resultou na geração de modelos que serão avaliados posteriormente. Com o uso dessas estratégias, o processo de aprendizado do algoritmo torna-se mais fácil e o desempenho pode ser otimizado.

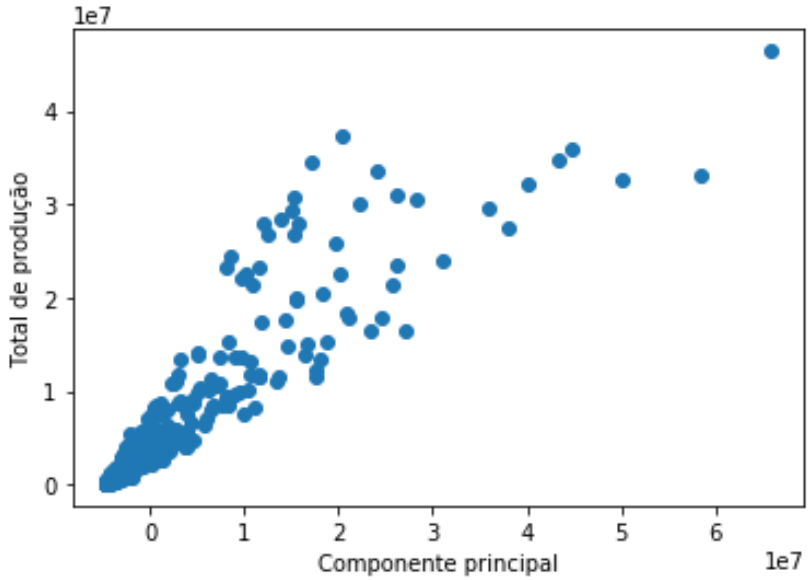
4.2.3 Análise dos Dados

A fase da análise tem por objetivo explorar os dados e encontrar padrões e correlações entre eles a fim de extrair informações relevantes que ajudam no entendimento do negócio. As informações adquiridas através dos formulários aplicados durante a oficina citada na Subseção 4.2.1 foram utilizadas para a análise exploratória de dados para identificação de correlações, padrões e tendências.

Para se ter uma análise da estrutura dos dados trabalhados nos experimentos, foi feito o uso do PCA, já definido no Capítulo 2. Esta abordagem ajuda a visualizar e entender melhor o comportamento dos dados, reduzindo a dimensionalidade do vetor de entradas à 1 dimensão. O PCA consiste basicamente em extrair informações de um conjunto de atributos inter-correlacionados e reduzir o tamanho do conjunto, mantendo apenas as informações importantes. Essa técnica foi utilizada com o intuito de auxiliar, durante esta fase, a perceber o comportamento dos dados. Dessa forma, foi possível visualizar a dispersão dos dados a partir de um gráfico de duas dimensões em função da produção de mel, que é o objetivo da predição dos modelos gerados neste trabalho, tanto no cenário do *dataset Honey Production in the USA*, como mostra a Figura 9, como também no cenário do *dataset* da Embrapa, representado pela Figura 10.

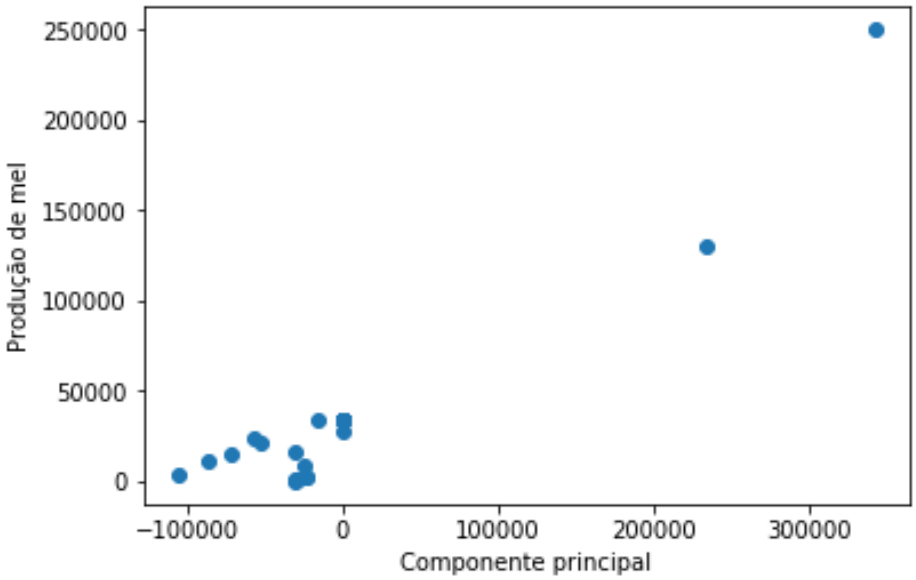
Foram realizados diferentes experimentos com o objetivo de encontrar o melhor modelo preditivo. Como já observado anteriormente, o segundo *dataset* tem um problema com a dimensionalidade. No caso, existe uma quantidade muito grande de atributos enquanto o número de amostras é muito pequeno. Com o objetivo de avaliar também modelos gerados com menores quantidades de atributos, buscou-se técnicas para ajudar a diminuir essa dimensionalidade. A abordagem utilizada neste trabalho

Figura 9 – Dispersão dos dados para o *dataset Honey Production in the USA*



Fonte: Elaborada pelo autor.

Figura 10 – Dispersão dos dados para o *dataset* da Embrapa



Fonte: Elaborada pelo autor.

para a seleção de características consiste na verificação do coeficiente de correlação de Pearson de todos os atributos com relação ao atributo alvo da predição (quantidade de mel produzida). Este coeficiente foi utilizado em trabalhos presentes na literatura, como em (IRANMANESH et al., 2013) para seleção de atributos na verificação de assinaturas *online* e em (MUJAHID; THIRUMALAI, 2017) para encontrar a relação entre os atributos de um adenoma a fim de prever um câncer.

A correlação de Pearson indica o grau de correlação entre duas variáveis além também da direção da correlação, sendo esta positiva ou negativa, podendo variar de -1 a 1. Uma correlação -1 indica uma correlação negativa muito forte entre as variáveis. Por exemplo, quanto mais uma cresce, mais a outra tende a diminuir. Já uma correlação de valor 1, indica uma correlação positiva muito forte entre as variáveis. Quanto mais próximo de 0, menor é a correlação entre as variáveis (FILHO et al., 2010). Com isso, foi possível realizar experimentos a partir de uma base com os atributos de maior correlação, seguindo a Equação 4.1, onde x_i e y_i são os valores dos atributos previsores e atributo alvo, respectivamente, assim como $\bar{x}$ e $\bar{y}$ são suas respectivas médias.

$$\rho = \frac{cov(x, y)}{\sqrt{var(x)\ var(y)}} \tag{4.1}$$

$$cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{4.2}$$

$$var(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.3}$$

$$var(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{4.4}$$

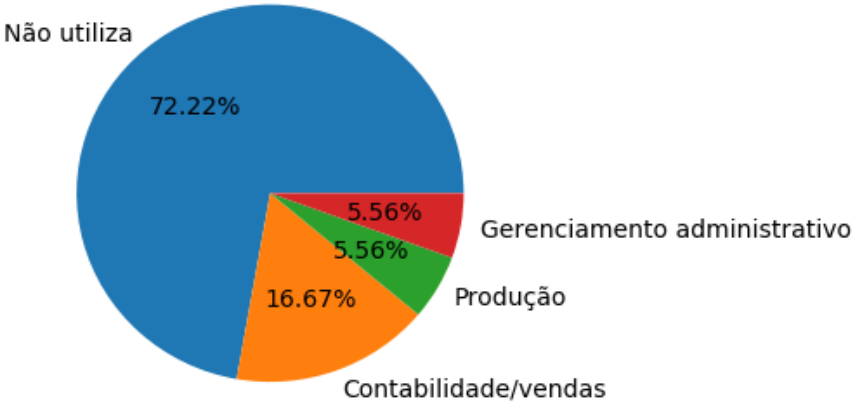
Para esse caso, foram considerados apenas os atributos que possuíam correlação acima de 80%. O *dataset* criado a partir dessa técnica contém 8 atributos identificados como importantes, mostrados na Tabela 3. O atributo em negrito é o atributo alvo da predição.

Através das análises realizadas, também foi possível confirmar a afirmação apresentada no início desse trabalho com relação ao uso de *software* para gerenciamento, como visto na Figura 11, o que só reforça a falta de adesão a recursos desse tipo. O gráfico mostra que mais de 70% dos exemplos coletados não utilizam nenhuma aplicação dessa natureza em nenhum tipo de gerenciamento.

Tabela 3 – *Dataset* 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.

Figura 11 – Análise do uso de programas de gerenciamento



Fonte: Elaborada pelo autor.

4.2.4 Modelagem e Avaliação

Após a execução das etapas anteriores, neste trabalho fez-se o uso de diferentes algoritmos de AM presentes na literatura, discutidos na Seção 2.3, para a tarefa de regressão, com o intuito de criar e comparar modelos preditivos gerados a partir de experimentos realizados em um ambiente controlado.

A estimativa de produção de mel a partir de um conjunto prévio de informações consiste em um problema de regressão, onde o atributo a ser predito é um valor contínuo, ou seja, uma tarefa preditiva em que o tipo de aprendizado é supervisionado. Isso quer dizer que os rótulos ou valores preditos para um novo exemplo se dão com base nos atributos já rotulados no *dataset*.

Tendo o *dataset* preparado, é preciso separá-lo de forma a criar um conjunto

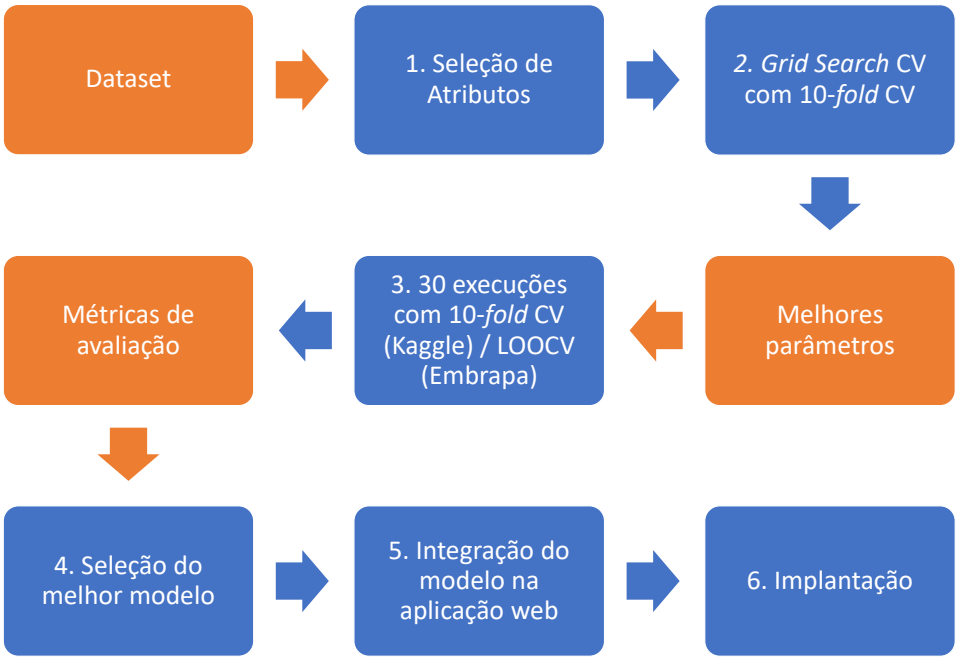
com os atributos previsores e um conjunto contendo o atributo alvo da predição. No caso de ambos os *datasets* utilizados, o objetivo da predição foi a quantidade de mel produzida. Logo, para o primeiro cenário, o atributo alvo é o *totalprod*, enquanto que no segundo cenário, o resultado da predição é o atributo C492016MEL_P, que corresponde à quantidade produzida referente a 2016, último ano informado nos questionários.

Para um melhor processamento de alguns algoritmos de AM, os dados necessitam estar distribuídos dentro de uma mesma escala. Para isso, os dados foram escalonados em ambos *datasets*, seguindo o método *StandardScaler* do *scikit-learn*. Este método consiste em subtrair a media x_{mean} dos valores de cada atributo x_i e escalonar dividindo pelo desvio padrão x_{std}. Cada valor é escalonado conforme a Expressão 4.5.

$$x_{scaled} = \frac{x_i - x_{mean}}{x_{std}} \tag{4.5}$$

A Figura 12 mostra o conjunto o *pipeline* utilizado durante os experimentos de AM.

Figura 12 – *Pipeline* dos experimentos de AM realizados



Fonte: Elaborada pelo autor.

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 AM utilizado durante o treinamento. Com esse intuito,

fez-se uso da técnica *GridSearchCV*, definida no Seção 2.5. O uso dessa técnica implica em selecionar manualmente uma *grid* de parâmetros correspondentes a cada algoritmo. Esses parâmetros selecionados são submetidos a cada algoritmo através do uso de uma *cross-validation* com o número de partições igual a 10 (parâmetro comumente usado na literatura em experimentos de AM), utilizando todo o *dataset* para realizar o treinamento e retornando os melhores parâmetros encontrados.

Os mesmos algoritmos foram utilizados para os experimentos realizados em ambos os *datasets*, portanto, seguem a mesma *grid* de parâmetros. A *grid* de parâmetros foi definida da seguinte forma:

- **Regressão Linear Múltipla**: *fit_intercept* e *normalize*;
- **Decision Tree**: *criterion, splitter* e *random_state=0*;
- **Random Forest**: *criterion, n_estimators* e *random_state=0*;
- **MLP**: *activation, solver, hidden_layers_sizes* e *random_state=0*;
- **SVR**: *kernel* e *degree*.

A validação das técnicas de AM em tarefas de regressão, geralmente envolve a realização de experimentos controlados, como forma de demonstrar a efetividade do desempenho de suas predições (FACELI et al., 2011). Para a execução desses experimentos, os *datasets* são divididos em conjuntos de treinamento e de teste. O modelo irá aprender o comportamento da base de dados passada no conjunto de treinamento e então submeter aos testes utilizando o conjunto separado para teste.

4.2.4.1 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 em conjuntos de treinamento e teste utilizando a técnica *K-fold Cross-validation*, explicada na Seção 2.6.1, como mostra na etapa 3 da Figura 12. Essa técnica foi escolhida pelo fato do *dataset* possuir um tamanho considerável de amostras. Foi utilizado, portanto, $K = 10$, sendo K o número de partições em que o conjunto foi dividido.

Os experimentos foram executados em 30 rodadas para cada algoritmo já com os hiperparâmetros selecionados a fim de garantir a veracidade dos resultados (JAIN, 1991). Em cada uma das rodadas, o *dataset* foi dividido em 10 subconjuntos mutuamente exclusivos e de mesmo tamanho, e executado 10 vezes em cada uma das 30 rodadas, sendo sempre escolhido um subconjunto para ser usado como teste. Dentro

de cada uma dessas rodadas, foi feito o treinamento utilizando um dos algoritmos de AM para ser o 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, explicadas na Seção 2.6, para as partições de dados geradas aleatoriamente. Ao final das 30 execuções foi verificada a média de todos os valores das métricas de avaliação gerados a partir da utilização da técnica *K-fold Cross-validation*, bem como o desvio padrão entre eles, dado pela Equação 4.6, onde x_i é o elemento na lista das métricas de avaliação registradas, $\bar{x}$ a média desses elementos e n o número de elementos.

$$x_{std} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}} \tag{4.6}$$

Além disso, também foi calculado o intervalo de confiança desses resultados para um nível de confiança de 95%. O intervalo de confiança consiste em representar uma estimativa mais confiável, mostrando o quanto o resultado da média pode variar. Na Equação 4.7, é dado o cálculo do intervalo de confiança de 95% para mais e para menos, onde $\bar{x}$ é a média das métricas de avaliação, x_{std} é o desvio padrão calculado e n o tamanho da lista de resultados. Para um nível de confiança de 95%, as médias cairão sempre dentro de $\pm 1,96\ (x_{std}/\sqrt{n})$ da média.

$$(\bar{x} - 1,96\ (\frac{x_{std}}{\sqrt{n}}); \bar{x} + 1,96\ (\frac{x_{std}}{\sqrt{n}})) \tag{4.7}$$

A partir dos resultados dessas métricas, é possível verificar o modelo mais bem avaliado, conforme a etapa 4 da Figura 12.

4.2.4.2 Experimentos com os Datasets construídos a partir dos questionários

Para esse cenário, os experimentos iniciaram utilizando o *dataset* completo formado a partir dos questionários adquiridos com a Embrapa Amazônia Oriental para verificação do resultado da predição. Como logo de início o resultado se mostrou péssimo, foi percebido que esse *dataset* tinha um problema de dimensionalidade, ou seja, uma quantidade muito grande de atributos irrelevantes que puderam ser melhorados, transformando-o em outros dois *datasets* contendo apenas parcelas de atributos. Os experimentos mostrados a partir daqui correspondem a utilização desses dois *datasets* gerados no intuito de encontrar o melhor modelo preditivo.

O primeiro deles foi gerado a partir da percepção do negócio por parte do autor juntamente com o orientador deste trabalho. A partir desse entendimento, foram

considerados arbitrariamente apenas alguns atributos que pudessem interferir diretamente no objetivo da predição e compor a primeira versão da PoC. Sendo assim, o primeiro *dataset* foi composto pelos atributos presentes na Tabela 4. O atributo em negrito é o alvo da predição.

Tabela 4 – *Dataset* 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.

O segundo *dataset* gerado foi obtido a partir da verificação da correlação entre os atributos, como explicado na Seção 4.2.3. O *dataset* consiste de atributos com correlação acima de 80% para com o atributo alvo e pode ser verificado na Tabela 3, também mostrada durante a Seção 4.2.3. Essas abordagens utilizadas para seleção de atributos para formar os diferentes *datasets* utilizados nos experimentos correspondem à etapa 1 da Figura 12.

No que diz respeito à etapa 3 da Figura 12, a técnica aplicada para dividir o *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. Essa técnica consiste em dividir a base deixando sempre apenas um exemplo para formar o conjunto de teste. Esse *dataset* era formado de 23 exemplos, como já citado, logo, os experimentos foram executados 23 vezes, onde a cada iteração, 22 exemplos foram usados para treinamento e 1 para teste. Dentro de cada iteração, o algoritmo de AM usado em cada experimento 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. Esses resultados ajudam a selecionar o melhor modelo, como é mostrado na etapa 4 da Figura 12.

Posteriormente, dentre os modelos mais bem avaliados nos diferentes *datasets*, será integrado junto ao sistema *Web* o modelo que mais se adéqua para o problema em questão e posteriormente implantado, representado as etapas 5 e 6, respectivamente, da Figura 12.

5 RESULTADOS

No Capítulo 4, foi apresentada a proposta deste trabalho a fim de gerar uma ferramenta para gestão apícola contando com um módulo de inteligência baseado em técnicas de AM para predição de produção de mel. Este capítulo aborda os resultados obtidos com o desenvolvimento da solução proposta, apresentando a comparação dos resultados obtidos a partir de métricas de avaliação dos modelos gerados, bem como as funcionalidades obtidas com o desenvolvimento do sistema *Web* proposto.

5.1 Avaliação de Desempenho dos Modelos de Predição

Os primeiros resultados apresentados consistem de uma abordagem comparativa do uso de diferentes modelos preditivos que satisfazem o problema em questão. A análise comparativa dos modelos foi conquistada a partir do uso de métricas de avaliação presentes na literatura, como forma de validar seus desempenhos a partir de experimentos controlados utilizando algoritmos de AM para a tarefa de regressão.

Como já citado durante o Capítulo 4, foram feitos diferentes experimentos com múltiplos *datasets* a fim de verificar a possibilidade de gerar modelos com eficiência preditiva para a produção de mel. Nesses resultados, será comparado o uso dos cinco algoritmos estudados durante a realização deste trabalho e explicados na Seção 2.3 para a geração dos modelos conquistados.

5.1.1 Avaliação no Dataset Honey Production in the USA

Primeiramente, são exibidos na Tabela 5 os resultados dos desempenhos 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.

Os resultados seguem as métricas de avaliação R^2 *Score*, que mostra o quão bem o modelo está se ajustando aos dados, e o cálculo do erro realizado a partir da métrica RMSE, que penaliza erros maiores e entrega o resultado na escala padrão dos dados. A métrica RMSE indica o quanto o modelo erra para a predição de um novo valor. 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

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

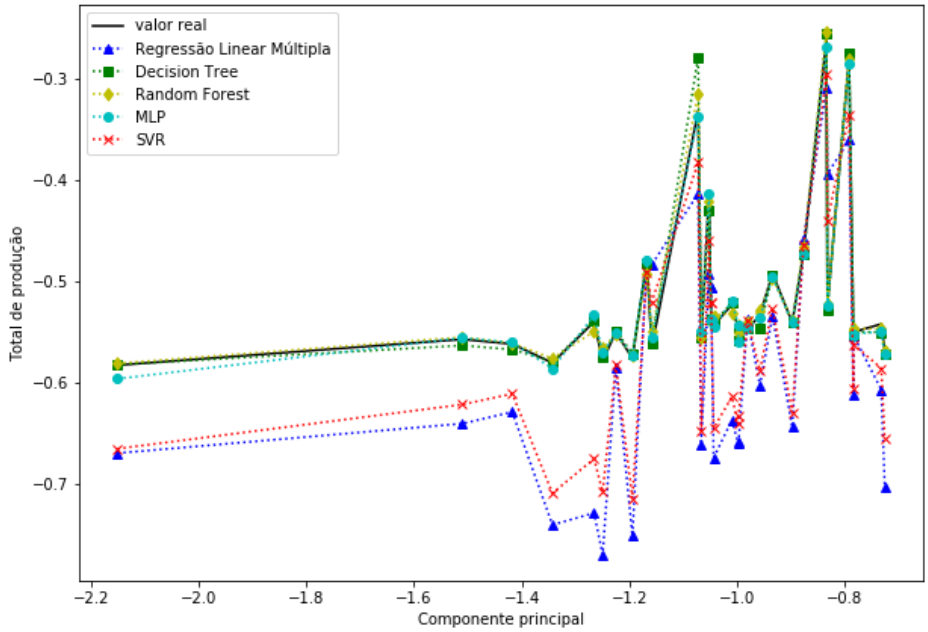
	R^2 **Score**	**Desv. Padrão do** R^2	**Int. Confiança do** R^2	**RMSE**	**Desv. Padrão da RMSE**	**Int. Confiança da RMSE**
Regressão Linear Múltipla	0,9523	0,0015982	(0,9517; 0,9529)	1.382.690	6.039,84	(1.380.526; 1.384.849)
Decision Tree	0,9624	0,0016880	(0,9618; 0,9630)	1.245.690	25.653,2	(1.236.513;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

realizadas. Porém, o modelo teve ainda um erro de cerca de 1.382.690 libras, onde o maior valor registrado no *dataset* é 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, errando cerca de 100.812 libras, que pode ser considerado baixo, quando comparado aos maiores valores presentes no *dataset*.

Ao usar o *K-Fold Cross-validation* com $K = 10$, o *dataset* de 626 exemplos foi dividido em 10 partições para cada uma das 30 rodadas. Em cada rodada, 62 exemplos são destinados para teste e o restante para treinamento. As Figuras 13 e 14 demonstram o gráfico dos resultados das predições na última rodada de experimentos utilizando o conjunto da última partição de teste. Como são muitos dados, resolveu-se dividir a leitura do gráfico das predições em duas partes para uma melhor visualização.

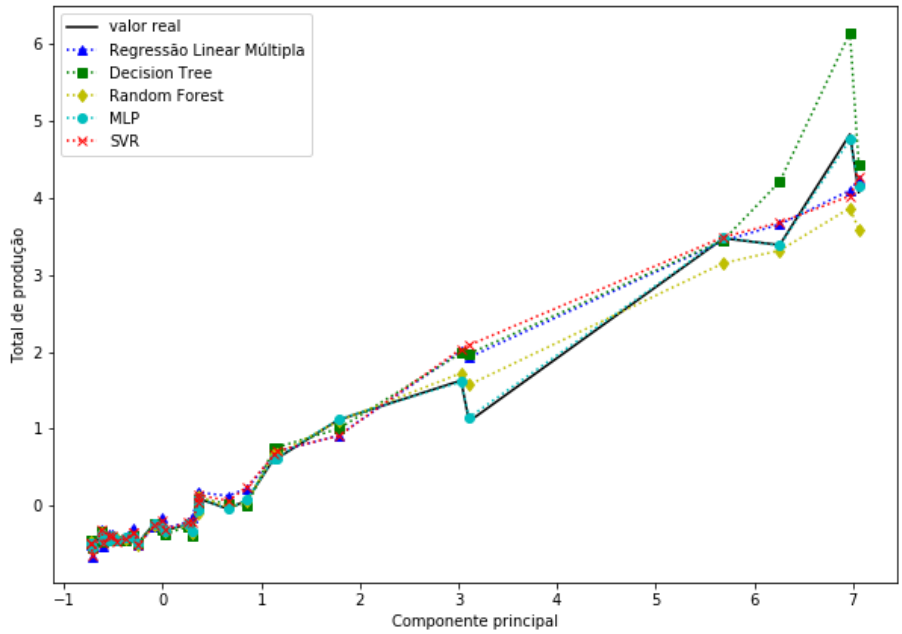
A partir do resultado de desempenho de forma gráfica, é possível perceber um melhor comportamento do modelo utilizando MLP, quase que acompanhando o valor real, alvo da predição. Os dados estão sendo exibidos com os valores escalonados a fim de garantir uma melhor visualização de como suas predições se comportaram.

Figura 13 – Desempenho dos modelos - primeiros 31 exemplos



Fonte: Elaborada pelo autor.

Figura 14 – Desempenho dos modelos - últimos 31 exemplos



Fonte: Elaborada pelo autor.

5.1.2 Avaliação nos Datasets Gerados dos Questionários

Os primeiros experimentos utilizando o *dataset* completo, com todos os atributos, feito a partir dos questionários conseguidos com a Embrapa Amazônia Oriental, não tiveram bons resultados, obtendo erros muito grandes. Após a conclusão de que haveria a necessidade de fazer uma seleção de atributos, foram criados outros *datasets* contendo parcelas diferentes de atributos para realização dos experimentos. O primeiro deles avaliado aqui, consiste de atributos escolhidos arbitrariamente pelos autores de acordo com a compreensão do negócio. O *dataset* gerado pode ser conferido através da Tabela 4, na Seção 4.2.4.2. Os resultados dos experimentos envolvendo o *dataset* em questão são apresentados na Tabela 6.

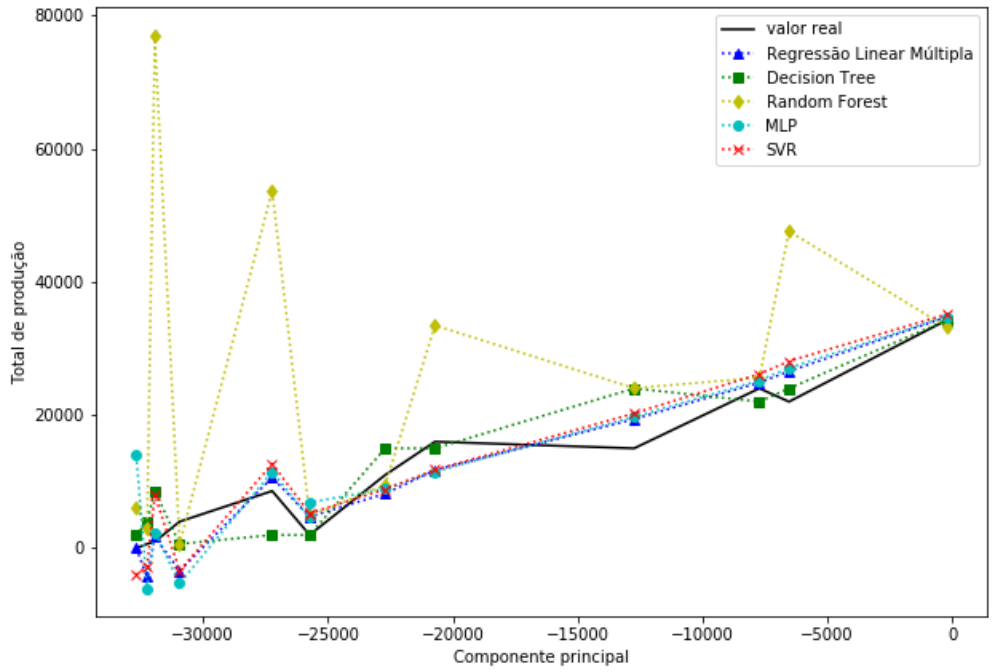
Tabela 6 – Resultados segundo as métricas de avaliação para o *dataset* com atributos escolhidos arbitrariamente

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

A partir dos resultados das métricas de avaliação R^2 *Score* e RMSE, foi possível inferir um melhor desempenho do modelo gerado utilizando o algoritmo Regressão Linear Múltipla. Os resultados mostram um *Score* de aproximadamente 99% do modelo, que da definição de R^2 na Seção 2.6.2, temos que esse valor representa a porcentagem de variação na resposta que é explicada pelo modelo, ou seja, indica o quanto o modelo consegue explicar os valores observados. 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 testados, obtendo um *Score* de 4% e a maior média de erro de 51.640,9 kg. O desempenho desse modelo seguiu acompanhado do 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 avaliado usou MLP com a função de ativação *logistic*, uma camada oculta com 100 neurônios e novamente, se tratando de bases pequena, o *solver* utilizado foi o lbfgs. O modelo obteve um *Score* de cerca de 35% e uma média de erro registrada 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 no caso dos experimentos envolvendo o *dataset* formado a partir dos questionários 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. Assim como no caso do *dataset* anterior, como alguns dados ficaram expostos de maneira muito próxima, dificultando a visualização, o gráfico contendo o desempenho da predição de cada modelo foi divido em dois. O primeiro, presente na Figura 15, contém os 12 primeiros exemplos e na Figura 16, são mostrados os 11 exemplos restantes.

Figura 15 – Desempenho dos modelos - primeiros 12 exemplos

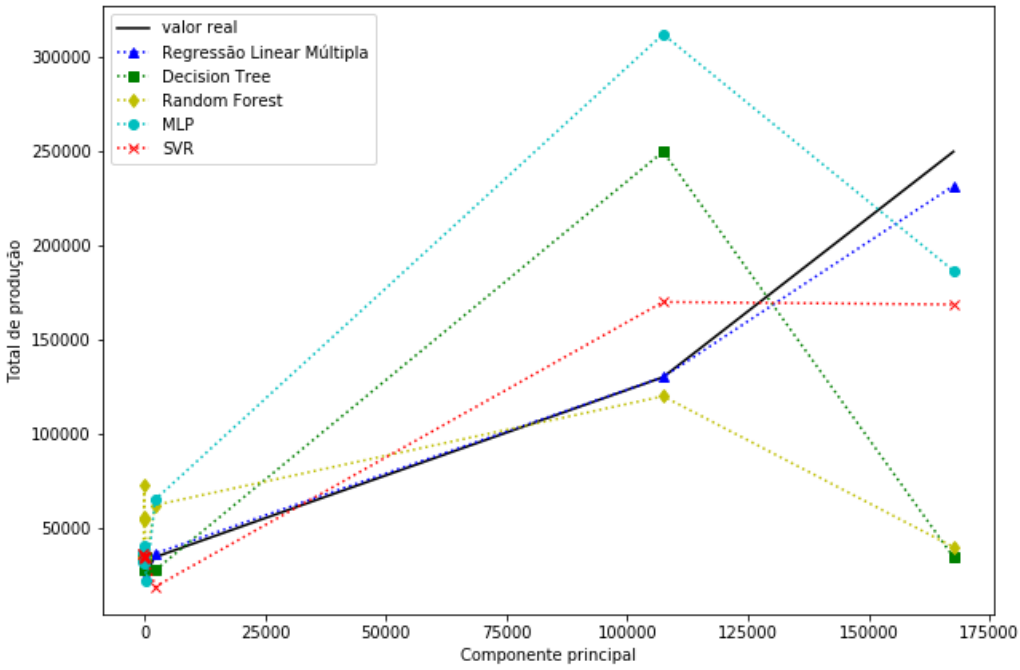


Fonte: Elaborada pelo autor.

Através dos gráficos, é possível conferir que o modelo mais bem avaliado (Regressão Linear Múltipla) teve um desempenho bem próximo da reta traçada pelos valores reais.

Com o intuito de melhorar ainda mais os resultados de desempenho, foram avaliados os modelos utilizados em experimentos a partir de um novo *dataset* 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

Figura 16 – Desempenho dos modelos - últimos 11 exemplos



Fonte: Elaborada pelo autor.

escolhidos foram os de correlação acima de 80%, como já explicado durante a Seção 4.2.3. 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%

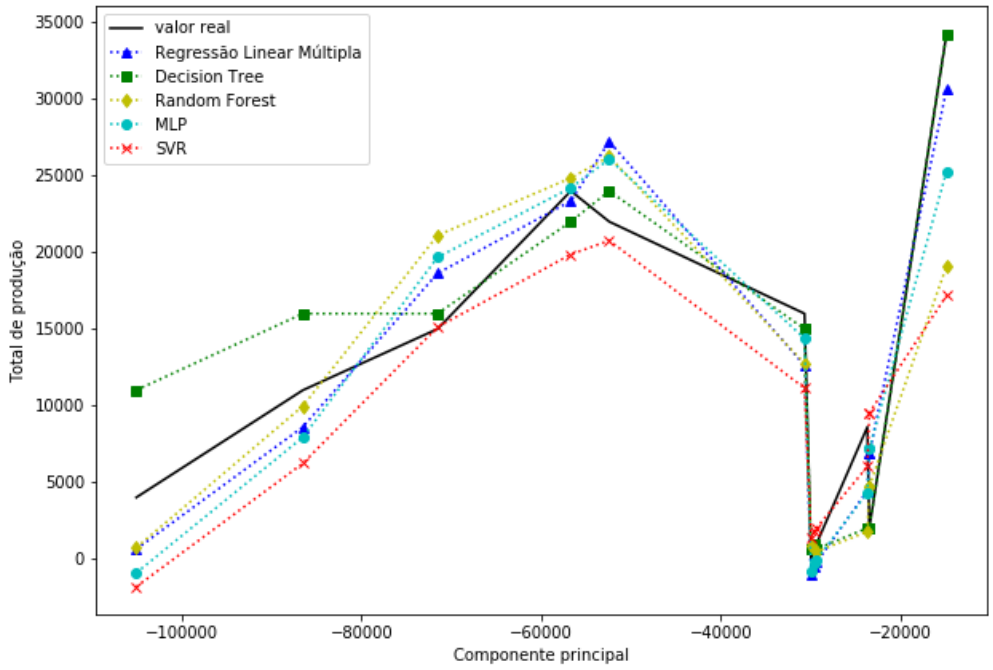
	R^2 **Score**	**RMSE**
Regressão Linear Múltipla	**0,9963**	**3.176,1**
Decision Tree	0,6308	32.133,2
Random Forest	0,5514	35.420
MLP	0,9026	16.496,9
SVR	0,9058	16.225,6

Como pode ser visto, 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 modelo que usa o algoritmo 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 resul-

tados 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 17 e a Figura 18 mostram o gráfico de desempenho dos modelos conforme os resultados da predição, divididos em duas partes para uma melhor visualização.

Figura 17 – Desempenho dos modelos - primeiros 12 exemplos



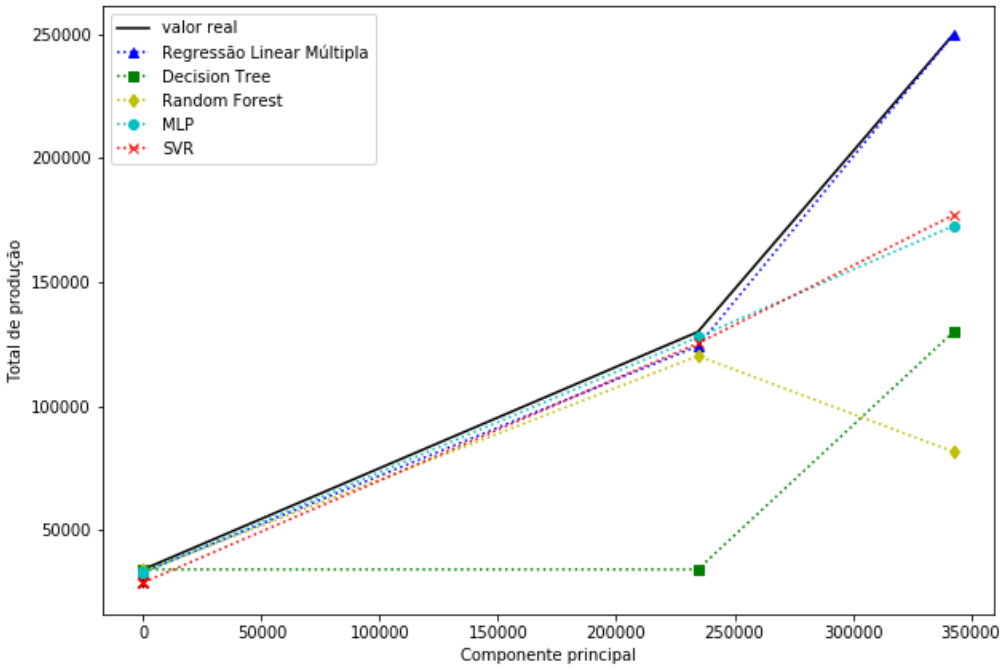
Fonte: Elaborada pelo autor.

Então, a partir dos experimentos realizados e avaliação dos modelos utilizados, foi possível perceber que o modelo utilizando Regressão Linear Múltipla é o que se ajusta melhor ao problema abordado por este trabalho, obtendo os melhores resultados de desempenho. Os resultados da avaliação preditiva permitem a escolha de um modelo para ser inserido na ferramenta *Web*. Com esse modelo, o usuário é capaz de analisar as características de suas propriedades e seus respectivos rendimentos previstos com base nas informações passadas.

5.2 Sistema *Web*

Os resultados do sistema proposto contam com o desenvolvimento das funcionalidades explicadas durante a Seção 4.1.

Figura 18 – Desempenho dos modelos - últimos 11 exemplos

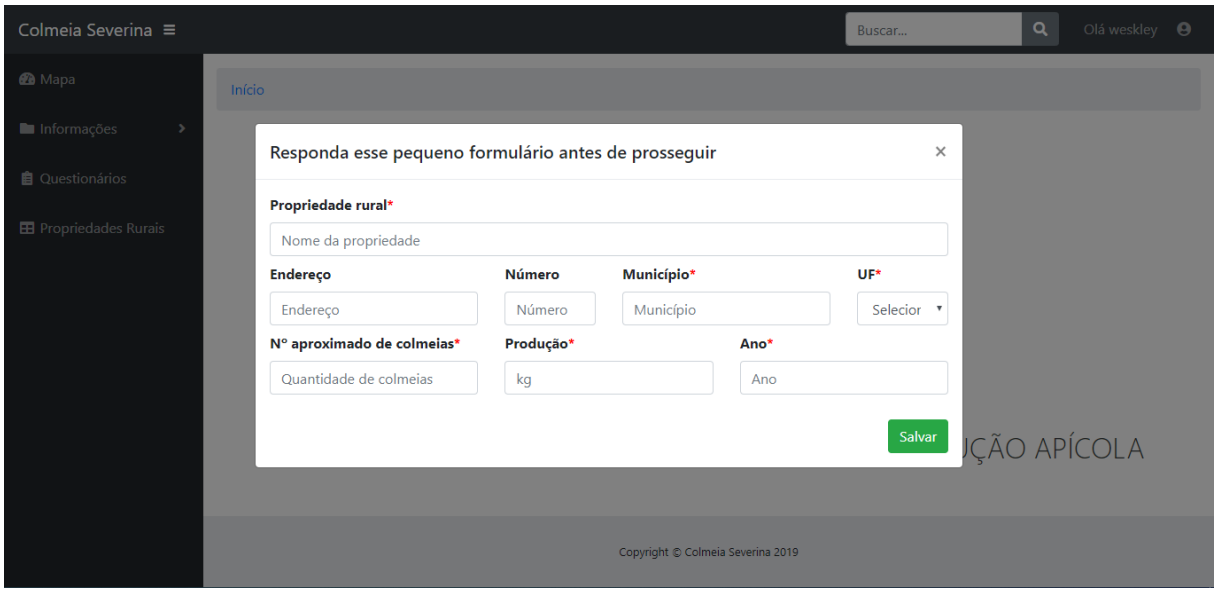


Fonte: Elaborada pelo autor.

O usuário, ao se cadastrar no sistema, realiza o *login* automaticamente e é direcionado à uma tela com um formulário a respeito de características sobre sua propriedade, como mostrado na Figura 19. As informações fornecidas são adicionadas ao banco de dados. Enquanto o usuário não fornecê-las ou não tiver nenhuma propriedade cadastrada em sua conta, continuará sendo notificada a necessidade de preenchimento do formulário ao se logar.

O usuário cadastrado só terá acesso às demais funcionalidades do sistema após confirmação da sua conta junto ao e-mail cadastrado. No momento do cadastro, é enviado um e-mail ao usuário contendo um *link* composto de um *token* que se encarrega de garantir a confirmação daquele usuário e o redireciona para a sua conta no sistema. Com a confirmação de sua conta, o usuário passa a ter acesso ao restante das funcionalidades que antes estavam limitadas. O usuário apicultor/gestor consegue ter acesso apenas às informações relacionadas à sua conta, então é possível gerenciar suas propriedades (Figura 20), assim como seus apiários (Figura 21) e, consequentemente, suas colmeias (Figura 22). Ainda assim, o usuário apicultor/gestor consegue ter acesso à ferramenta de geolocalização de todas as propriedades cadastradas em sua região dispostas em um mapa através de marcadores, como mostra a

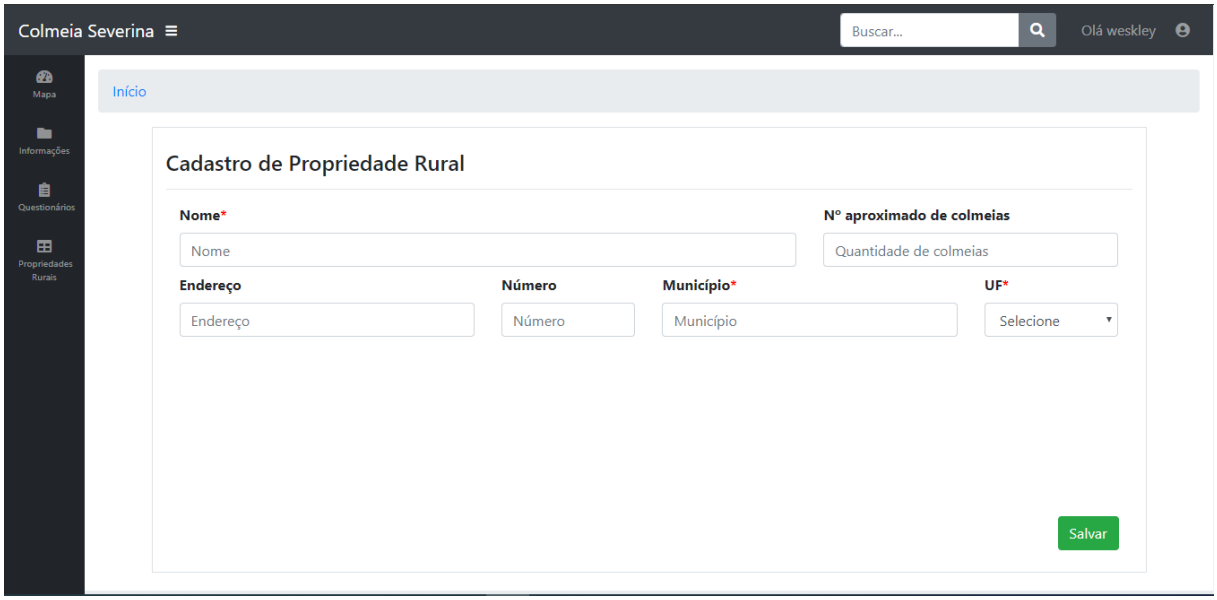
Figura 19 – Tela de formulário inicial



Fonte: Elaborada pelo autor.

Figura 23.

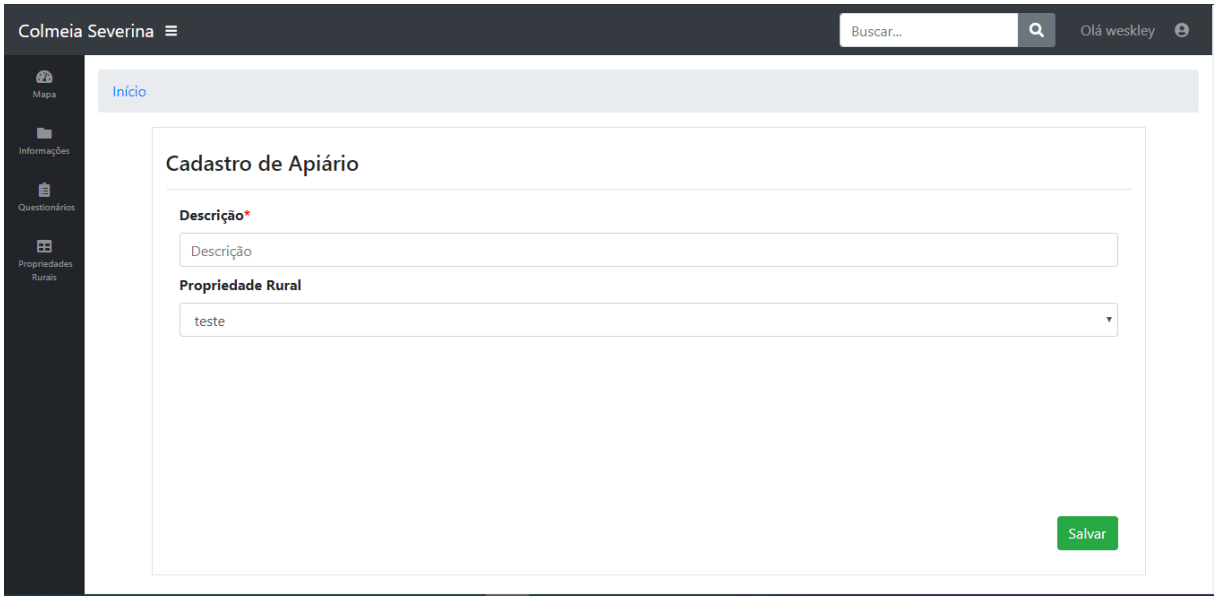
Figura 20 – Tela de cadastro de propriedade



Fonte: Elaborada pelo autor.

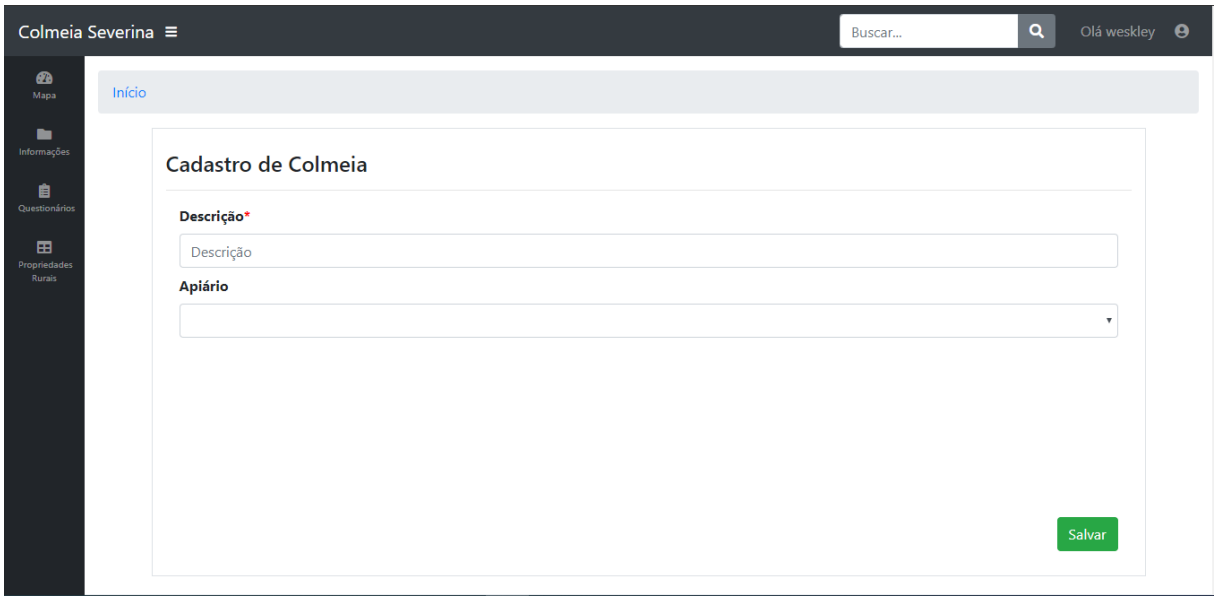
O administrador do sistema tem acesso aos dados sobre todos os usuários e as informações relacionadas a eles, mantendo o controle sobre o cadastro dos usuários do sistema e suas respectivas propriedades.

Figura 21 – Tela de cadastro de apiário



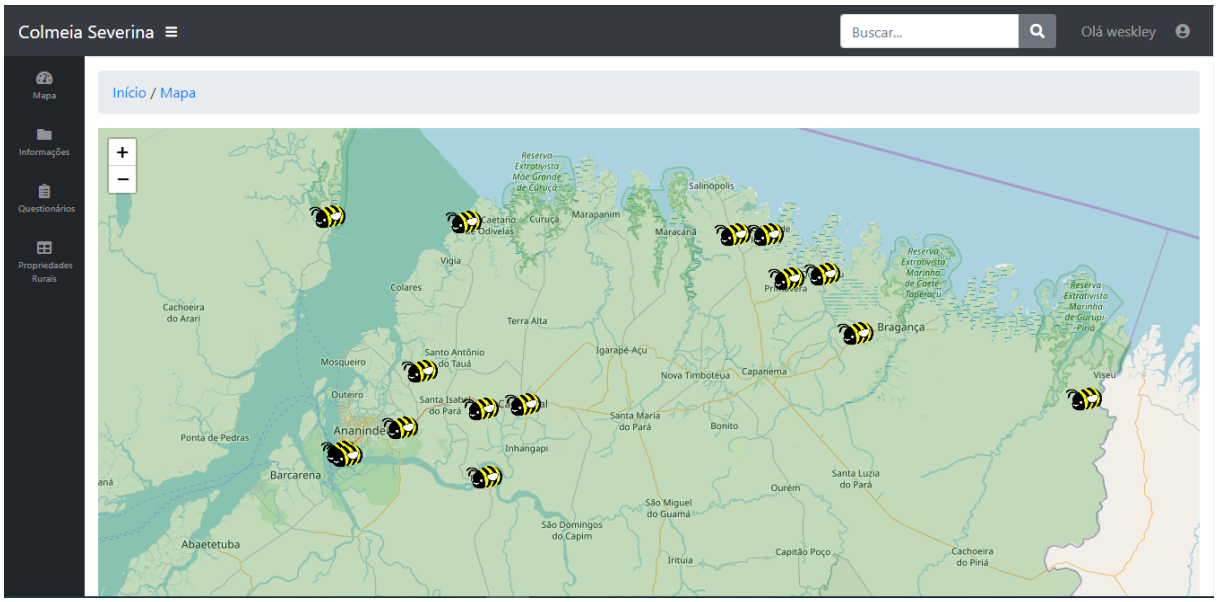
Fonte: Elaborada pelo autor.

Figura 22 – Tela de cadastro de colmeia



Fonte: Elaborada pelo autor.

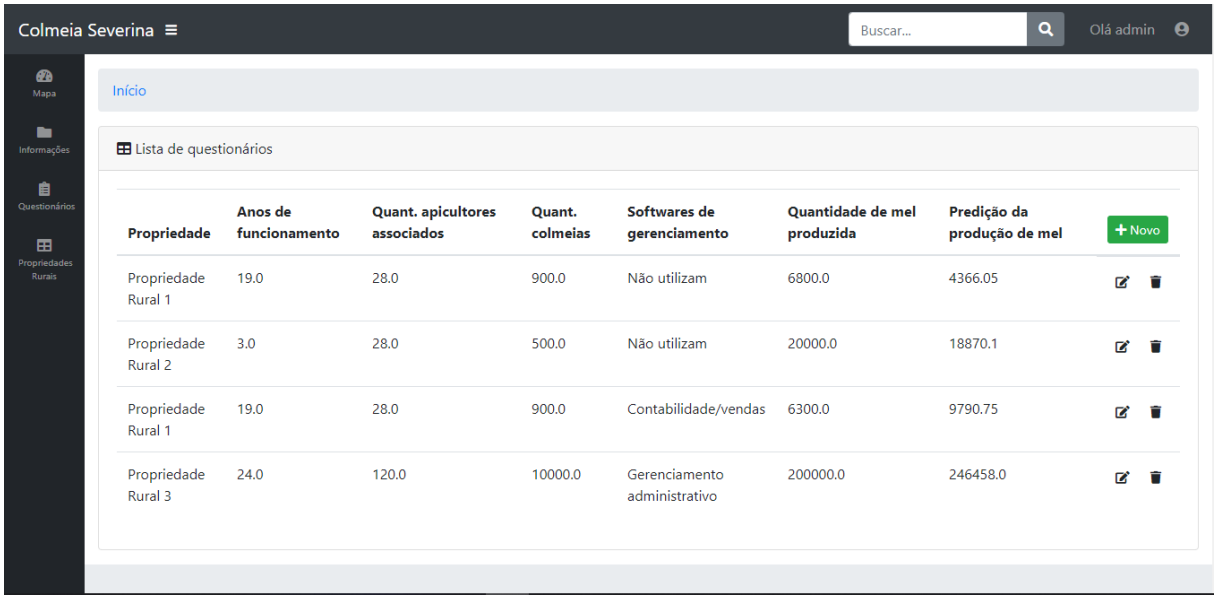
Figura 23 – Tela de georreferenciamento do sistema



Fonte: Elaborada pelo autor.

Em se tratando do módulo de predição integrado ao sistemas *Web*, os resultados conquistados mostram uma tela onde são passadas as informações necessárias para o modelo de AM realizar a predição por meio de um questionário. 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 24.

Figura 24 – Tela contendo a realização da predição de 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: Elaborada pelo autor.

6 CONCLUSÃO E TRABALHOS FUTUROS

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, como citado no decorrer deste trabalho, afeta diretamente nos níveis de produção. O pequeno produtor rural, muitas vezes, não dispõe de nenhum tipo de tecnologia para gerenciar seus recursos, e auxiliá-lo de forma sábia nas tomadas de decisões.

Levando em consideração essa lacuna existente no cenário brasileiro, neste trabalho foi apresentada uma ferramenta *Web* para gerenciamento apícola e predição de produção de mel através de uma *engine* de inteligência baseada em técnicas de AM, tendo em vista a carência apresentada no setor apícola quanto ao uso de tecnologias nos processos de gestão e produção. Este trabalho contou com a realização 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 AM voltados para a tarefa de regressão, para realizar predições a respeito da produção de mel. 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 dentro do contexto trabalhado. 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 criado 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 contam ainda com o desenvolvimento de um sistema *Web* como prova de conceito deste trabalho. 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 isso, o desenvolvimento de um sistema *Web* possibilita 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* é responsável por manter 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.

A partir de tudo que foi exposto neste trabalho, ficam como sugestões de trabalhos futuros, a possibilidade da aplicação de técnicas de seleção de atributos presentes na literatura, de forma exaustiva, a fim de avaliar suas contribuições. Além disso, para contrapor uma das maiores dificuldades encontradas neste trabalho, 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. O módulo *Web* também continua seu desenvolvimento, podendo ser aprimorado através de *dashboards* que facilitem o acompanhamento e a tomada de decisão através da implementação de informações gráficas, por exemplo.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. Citado na página 28.

AMARAL, F. *Introdução à Ciência de Dados: mineração de dados e big data*. [S.l.]: Alta Books Editora, 2016. Citado na página 19.

BRASIL. Portaria nº 162, de 24 de abril de 2014. estabelece as rotas de integração nacional como a estratégia de inclusão produtiva e desenvolvimento regional do ministério da integração nacional, e institui o comitê gestor das rotas. *Diário Oficial da União*, Poder Executivo Brasília, DF, n. 162, 2014. Citado na página 16.

CERQUEIRA, A.; FIGUEIREDO, R. A. Percepção ambiental de apicultores: Desafios do atual cenário apícola no interior de são paulo. *Acta Brasiliensis*, v. 1, n. 3, p. 17–21, 2017. Citado na página 14.

DUTRA, T. F. S. *Beehiveior-Sistema de monitoramento e controle de colmeias de produção apícola*. Dissertação (Mestrado) — Brasil, 2016. Citado na página 16.

EMBRAPA. *Embrapa Amazônia Oriental - Oficina prepara inclusão da Amazônia na Rota do Mel*. 2017. https://www.embrapa.br/busca-de-noticias/-/noticia/21352318/oficina-prepara-inclusao-da-amazonia-na-rota-do-mel. Acessado em 22-01-2019. Citado na página 16.

EMBRAPA. *Secretaria de Estado de Ciência, Tecnologia e Educação Profissional e Tenológica - Governo do Estado do Pará. Rota do Mel. Rota do Mel*. 2017. http://www.sectet.pa.gov.br/noticias/rota-do-mel. Acessado em 22-01-2019. Citado na página 16.

EMBRAPA. *SUDAM - Superintendência do Desenvolvimento da Amazônia: Oficina debate Rota do Mel na Região Norte*. 2017. http://www.sudam.gov.br/index.php/ouvidoria/17-ultimas-noticias/1041-oficina-discute-rota-do-mel-na-regiao-norte. Acessado em 22-01-2019. Citado na página 16.

FACELI, K. et al. *Inteligência Artificial: Uma Arbodagem de Aprendizagem de Máquina*. [S.l.]: LTC, 2011. Citado 12 vezes nas páginas 14, 19, 20, 21, 23, 25, 26, 27, 28, 29, 41 e 47.

FILHO, A. C. et al. Tamanho de amostra para estimação do coeficiente de correlação linear de pearson entre caracteres de milho. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 45, n. 12, p. 1363–1371, 2010. Citado na página 44.

GILIOLI, G. et al. Multi-dimensional modelling tools supporting decision-making for the beekeeping sector. *IFAC-PapersOnLine*, Elsevier, v. 51, n. 5, p. 144–149, 2018. Citado na página 31.

GRANDÓN, N. et al. Information system for improving local productivity and decision making in organic beekeeping. In: IEEE. *Automatica (ICA-ACCA), IEEE International Conference on*. [S.l.], 2016. p. 1–7. Citado 4 vezes nas páginas 15, 16, 32 e 33.

GRUS, J. *Data Science do zero: Primeiras regras com o Python*. [S.l.]: Alta Books Editora, 2018. Citado 2 vezes nas páginas 28 e 36.

HALL, M. A. Correlation-based feature selection for machine learning. University of Waikato Hamilton, 1999. Citado na página 55.

IBGE. *Produção da Pecuária Municipal*. 2016. https://biblioteca.ibge.gov.br/visualizacao/periodicos/84/ppm_2016_v44_br.pdf. Acessado em 19-01-2019. Citado na página 15.

IBGE. *Pesquisa da Pecuária Municipal*. 2018. https://sidra.ibge.gov.br/pesquisa/ppm/quadros/brasil/2018. Acessado em 19-01-2019. Citado na página 15.

IRANMANESH, V. et al. Online signature verification using neural network and pearson correlation features. In: IEEE. *2013 IEEE Conference on Open Systems (ICOS)*. [S.l.], 2013. p. 18–21. Citado na página 44.

JAIN, R. *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling.* [S.l.]: Wiley, 1991. I-XXVII, 1-685 p. (Wiley professional computing). ISBN 978-0-471-50336-1. Citado na página 47.

JOSHI, P. *Artificial intelligence with python*. [S.l.]: Packt Publishing Ltd, 2017. Citado na página 22.

KARADAS, K.; KADIRHANOGULLARI, I. H. Predicting honey production using data mining and artificial neural network algorithms in apiculture. *Pakistan Journal of Zoology*, AsiaNet Pakistan (Pvt) Ltd., v. 49, n. 5, 2017. Citado 3 vezes nas páginas 16, 33 e 34.

LUÍS, T. V. F. *Business Intelligence para apoio à gestão das listas de inscritos para cirurgia em Portugal continental*. Tese (Doutorado), 2014. Citado na página 19.

MARANHÃO, P. B. d. A. A. et al. Avaliação dos métodos de custeio na produção de mel: um estudo de caso no município de são joão do rio do peixe. Universidade Federal de Campina Grande, 2016. Citado na página 14.

MAROS, A. et al. Aprendizado de máquina para previsão do tempo de execução de aplicações spark. In: SBC. *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. [S.l.], 2019. p. 197–210. Citado na página 24.

MILES, J. R squared, adjusted r squared. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, 2014. Citado na página 29.

MUJAHID, A. K.; THIRUMALAI, C. Pearson correlation coefficient analysis (pcca) on adenoma carcinoma cancer. In: IEEE. *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. [S.l.], 2017. p. 492–495. Citado na página 44.

NORVIG, P.; RUSSELL, S. *Inteligência Artificial: Tradução da 3a Edição*. [S.l.]: Elsevier Brasil, 2014. Citado 2 vezes nas páginas 21 e 23.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 28.

PINTO, L. A. A. *Construção de aplicativo para o planejamento e gestão da produção apícola no Centro Paulista*. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado 4 vezes nas páginas 14, 15, 16 e 32.

PRAKASH, S.; SHARMA, A.; SAHU, S. S. Soil moisture prediction using machine learning. In: IEEE. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. [S.l.], 2018. p. 1–6. Citado 2 vezes nas páginas 22 e 27.

PROVOST, F.; FAWCETT, T. Data science para negócios. *Rio Janeiro: Alta Books*, 2016. Citado 2 vezes nas páginas 20 e 39.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. Citado 2 vezes nas páginas 19 e 20.

SILVA, R. A. da; SILVA, F. C. A.; GOMES, C. F. S. O uso do business intelligence em sistema de apoio a tomada de decisão estratégica. *Innovation, Technology and Management Journal*, v. 6, n. 1, p. 2780–2798, 2016. Citado 2 vezes nas páginas 14 e 19.

SOMMERVILLE, I. *Engenharia de Software: Tradução da 9a Edição*. [S.l.]: Pearson Education, Inc., 2011. Citado na página 35.

THAMIR, A.; POULIS, E. Business intelligence capabilities and implementation strategies. *International Journal of Global Business*, Global Strategic Management Inc, v. 8, n. 1, p. 34, 2015. Citado na página 19.

VIDAL, M. de F. *Desempenho da Apicultura Nordestina em Anos de Estiagem*. 2017. https://www.bnb.gov.br/documents/80223/2130269/apicultura_11_2017.pdf/. Acessado em 19-01-2019. Citado na página 15.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 2 vezes nas páginas 24 e 38.

Anexos

ANEXO A - Questionário Rota do Mel

Rota do Mel

CBA Confederação Brasileira de Apicultura

Embrapa

Rotas de INTEGRAÇÃO Nacional

Diagnóstico nacional: Rota de do Mel – Macrorregião Nordeste

QUESTIONÁRIO "ASSOCIAÇÕES/COOPERATIVAS/ FEDERAÇÕES/ÓRGÃOS"

PESQUISADOR : _____ **DATA: / /2017**

A – IDENTIFICAÇÃO DA ASSOCIAÇÃO/COOPERATIVA/FEDERAÇÃO/ÓRGÃO

1.Nome :_____

2.Endereço :_____ 3.Fone: _____

_____ 4.CEP:_____

5. Município :_____ 6. Estado: _____

7. Entrevistado : _____

8. Cargo/Função :_____

B – CARACTERIZAÇÃO DAS ASSOCIAÇÕES/COOPERATIVAS/FEDERAÇÃO/ÓRGÃO

09. Quantos anos de funcionamento a entidade possui ?_____

10. Qual o número total de entidades associadas?_____

11. Qual a área de jurisdição da entidade?

_____

_____

_____

12. Quantos apicultores estão associados à entidade? _____

13. Qual o número aproximado de colmeias? _____

Capacitação

14. Foi promovido algum treinamento técnico para o quadro de sócios?
1. Não **2.** Sim, apenas teórico **3.** Sim, teórico e prático

15. Se sim, qual o órgão responsável pelo treinamento:
1. SENAR **2.** SEBRAE **3.** Universidade _____
5. ATER **6.** EMBRAPA **7.** ONG _____
8. Outro _____

16. Qual(is) a(s) área(s) em que realizou o treinamento?
1. Qualidade e produção
2. Operação de Máquinas e Equipamentos
3. Vendas e Marketing
4. Educação básica sobre apicultura
5. Gestão (especificar)_____

6. Outros _____

17. Houve algum treinamento de Gestão para os dirigentes?
1. Não **2.** Sim, teórico e prático **3.** Sim, apenas teórico

18. Qual o órgão que realizou o treinamento?
1. SENAR **2.** SEBRAE **3.** Universidade _____
5. ATER **6.** EMBRAPA **7.** ONG _____
8. Outro _____

19. Em qual(is) outra(s) área(s) houve treinamento para dirigentes?
1. Qualidade e produção
2. Operação de Máquinas e Equipamentos
3. Vendas e Marketing
4. Educação básica sobre apicultura
5. Custo de Produção
6. Outros _____

Recursos humanos

20. Qual o tipo de mão-de-obra predominante empregada na atividade apícola na região de atuação da Associação/Cooperativa/Federação?
1. Familiar remunerada **2.** Familiar não remunerada
3. Contratada permanente **4.** Contratada temporária
5. Troca de serviço **6.** Não emprega
7. Outros

21. Se a mão-de-obra é **contratada**, qual a forma de pagamento?
1. Produto **2.** Dinheiro **3.** Produto e Dinheiro **4.** Outros

22. Quantas pessoas, em média, o associado contrata para a atividade apícola?
1. 1 a 3 **2.** 4 a 6 **3.** 7 a 10
4. 11 a 15 **5.** 16 a 20 **6.** Acima de 20

23. Qual o valor médio da diária (R$) paga na região no *manejo* das colmeias ? _____

24. Qual o valor médio da diária paga na região (R$) na extração do mel ?_____

25. Qual o principal problema em relação a recursos humanos que afeta a atividade apícola?
1. Baixa escolaridade dos apicultores
2. Falta de treinamento técnico dos apicultores
3. Falta de treinamento gerencial dos apicultores
4. Falta de conhecimento técnico da mão-de-obra contratada
5. Falta de assistência técnica
6. Resistência cultural para uso de tecnologia
7. Outro _____

Acesso ao crédito

26. Qual o tipo de investimento na atividade apícola da maioria dos associados?
1. Recurso próprio **2.** Financiamento **3.** Recurso próprio e Financiamento **4.** Doação

27. Qual é a origem do financiamento feito na atividade apícola ?
1. Crédito bancário direto ao consumidor
2. Empréstimo de pessoa jurídica
3. Empréstimo de pessoa física
4. Programas governamentais de apoio a produção
5. Associação/Cooperativa ao qual está vinculado

6. ONG's
7. Outro

28. Em geral, qual(is) o(s) objetivo(s) do **empréstimo bancário**?
1. Capital de giro **2.** Investimento **5.** Outro
3. Custeio **4.** Capital de giro e investimento

29. O financiamento incluiu pacote tecnológico?
(Acesso à inovação tecnológica e assistência técnica)
1. Não **2.** Sim

30. Houve dificuldade na aquisição do empréstimo bancário?
1. Não **2.** Sim

31. Qual a *principal* dificuldade na aquisição de empréstimo bancário?
1. Juros elevados **2.** Exigências de garantias reais
3. Prazo de pagamento curto **4.** Restrições cadastrais
5. Recurso de contrapartida **6.** Nenhuma
7. Demora na liberação dos recursos **8.** Outra _____

Inovações tecnológicas no processo produtivo

32. Os apicultores recebem algum tipo de assistência técnica?
1. Não **2.** Sim

33. Os empreendimentos recebem algum tipo de assistência gerencial?
1. Não **2.** Sim

34. Os empreendimentos ou apicultores utilizam programas ou aplicativos de gerenciamento?
1. Não utiliza **2.** Gerenciamento administrativo
3. Contabilidade/vendas **4.** Produção **5.** Todos

35. Quais os principais problemas em relação ao gerenciamento no campo?
1. Mão-de-obra qualificada **2.** Aquisição de insumos de qualidade
3. Contabilidade **4.** Falta de planejamento das atividades
5. Comercialização **6.** Falta de conhecimento sobre gestão
7. Outros: _____

36. Quais os principais problemas em relação ao gerenciamento administrativo?
1. Mão-de-obra qualificada **2.** Aquisição de insumos de qualidade
3. Contabilidade **4.** Gerenciamento financeiro
5. Comercialização **6.** Gerenciamento de estoque
7. Outros: _____

37. Os empreendimentos ou apicultores utilizam controle de qualidade na produção?
1. Em nenhuma etapa **2.** Em algumas etapas
3. Em todo o processo produtivo **4.** Só para produtos acabados

38. Os empreendimentos ou apicultores utilizam controle de qualidade em relação aos insumos ou componentes?
1. Não é realizado **2.** No recebimento de todas as entregas
3. No recebimento de alguns insumos **4.** No laboratório

39. Quais as normas técnicas utilizadas quanto ao processo de **beneficiamento** do produto?
1. Da série ISO **2.** Normas do Ministério da Agricultura
3. Certificadora orgânica **4.** Comércio Justo e Solidário
5. ABNT **6.** Normas próprias dos compradores
7. Outro: _____

40. Qual o *principal* obstáculo p/ implantação de programas de incremento (ou melhoria) da produtividade e da qualidade?
    1. Falta de orientação técnica **2.** Falta de recursos financeiras
    3. Nível de qualificação de pessoal **4.** Falta de informações
    5. Não há interesse ou motivação **6.** Outros

C – COMERCIALIZAÇÃO NA ASSOCIAÇÃO/COOPERATIVA/FEDERAÇÃO
Mercado e Abastecimento

41. Qual percentual das vendas é destinado ao mercado local? _____

42. Que percentual das vendas é destinado ao mercado regional? _____

43. Que percentual das vendas é destinado ao mercado nacional? _____

44. Que percentual das vendas é destinado ao mercado internacional? _____

45. Que percentual das vendas é destinado a venda para programas de governo? _____

46. Qual o *principal* tipo de cliente?
    1. Consumidor final **2.** Varejo
    3. Atacado **4.** Cooperativas
    5. Indústria de pequeno e médio porte **6.** Industria de grande porte
    7. Comércio exportador **8.** Atravessador
    9. Distribuidor

47. A capacidade de produção/vendas nos últimos 3 anos
    1. Continua a mesma **2.** Aumentou **3.** Reduziu

48. Qual o percentual do aumento ou da redução anual? _____

49. Quais os tipos de produtos e a quantidade comercializados?

Produtos / Ano	2014		2015		2016	
	Produzida	Comercial.	Produzida	Comercial.	Produzida	Comercial.
1. Mel (kg)						
2. Pólen apícola (kg)						
3. Própolis (kg)						
4. Geleia real (kg)						
5. Cera (kg)						
6. Rainha (und)						
7. Apitoxina (kg)						
8. Enxames (unidade)						

54. Qual o custo de aquisição (R$/kg) de compra do mel do apicultor? _____
    (1L ≈ 1,4 kg mel)

55. Qual o preço de venda (R$/kg) do mel a granel para Associações e cooperativas? _____

56. Qual o custo de aquisição (R$/kg) do mel fracionado (beneficiado) para Associações e cooperativas? _____

57. Qual o preço de venda (R$/kg) do mel fracionado (beneficiado) pelas Associações e cooperativas? _____

58. Qual a principal forma de comercialização do mel?
    1. "*In Natura*" a granel **2.** Composto **3.** "In Natura" fracionado

59. Qual o *principal* problema encontrado no comércio?
1. Preço baixo
2. Falta de comprador
3. Alta taxa de imposto
4. Falta de publicidade
5. Desconhecimento de alternativas de vendas
6. Falta de produção p/ atender o mercado consumidor de grande porte
7. A qualidade do produto não atende ao mercado consumidor
8. Outro _____

60. Qual o principal problema de acesso ao **mercado nacional**?
1.Desconhecimento de procedimentos administrativos
2. Excesso de burocracia
3. Exigências de normas técnicas
4. Falta de financiamento
5.Falta de contato com representações nacionais
6. Dificuldade de associar-se com parceiros nacionais
7.Falta de participação em feiras ou exposições nacionais
8. Exigências legais dos estados importadores
9. Insuficiência do volume de produto a ser exportado
10. Produto sem controle de qualidade compatível
11. Produto sem especificação adequada
12. Barreiras alfandegárias
13. Falta de publicidade
14. Embalagens ou acondicionamentos inadequados
15. Outro _____

Mercado Exterior

61. Realiza exportação?
1. Não
2. Sim

62. Qual o principal problema de acesso ao **mercado externo**?
1.Desconhecimento de procedimentos administrativos
2. Excesso de burocracia
3. Exigências de normas técnicas
4. Falta de financiamento
5.Falta de contato com representações estrangeiras
6. Dificuldade de associar-se com parceiros estrangeiros
7.Falta de participação em feiras ou exposições internacionais
8. Exigências legais dos países importadores
9. Insuficiência do volume de produto a ser exportado
10. Produto sem controle de qualidade compatível
11. Produto sem especificação adequada
12. Barreiras alfandegárias
13. Falta de publicidade
14. Embalagens ou acondicionamentos inadequados
15. Outro _____

63. Qual o *principal* fator que favorece a inserção no **mercado exterior**?
1. Preço
2. Qualidade do produto
3. Infraestrutura adequada
4. Tipo de produto
5. Participação em feiras e exposições
6. Canal de comercialização adequado

64. Qual a *principal* exigência do **mercado externo** sobre o produto?
1. Padrão de embalagem
2. Novas técnicas de processo (série ISO)
3. Exigências fitossanitárias
4. Exigências de proteção ambiental
5. Outros: _____

65. Como acompanha as tendências do **mercado externo**?
1. Não acompanha
2. Através de feiras e congressos/revistas técnicas
3. Visitas ao exterior através dos clientes
4. Através de fornecedores de equipamentos

66. Qual a *principal* dificuldade enfrentada no processo burocrático necessário para **exportação**?
    1. Fase de licenciamento junto ao registro de exportadores e importadores (REI) e Sistema Integrado do Comércio Exterior (SISCOMEX)
    2. Fase aduaneira (despacho e desembaraço da exportação) junto à Receita Federal
    3. Fase cambial (receber o dinheiro e fechar o câmbio) junto ao Banco Central

67. Qual o *principal* fator que influencia o preço do **produto exportado** ?
    1. Custo matéria-prima
    2. Custo mão-de-obra
    3. Impostos e taxas
    4. Custo de embalagem especial
    5. Custo de Transporte
    6. Despesas administrativas/despesas financeiras