A digitalização do campo e a democratização da ciência de dados: perspectivas para aplicação por produtores agropecuários

Leonardo de Oliveira Dresch[1]
Adriano Marcos Rodrigues Figueiredo[2]
Mayra Batista Bitencourt Fagundes[3]

Submissão: 30/10/2021 Aceitação: 28/01/2022

Resumo

O contexto da pandemia da Covid-19, entre as suas diversas consequências, acelerou o processo de digitalização do campo e estimulou a adoção da internet pelos produtores agropecuários. A ciência de dados (*data science*) a cada dia que passa se torna mais acessível e já começa a chegar nas mais diversas organizações, inclusive as relacionadas a agropecuária e ao agronegócio. O *paper* buscou discutir as potencialidades oriundas da digitalização do campo e da utilização da ciência de dados com poucos recursos pelos produtores agropecuários. Trata-se de um ensaio acadêmico, baseado privilegiadamente em pesquisa bibliográfica. Os principais resultados apontam que os produtores podem se beneficiar de inúmeras formas da digitalização e da ciência de dados, entretanto o principal desafio é o aumento da abrangência do acesso à internet e o desenvolvimento da cultura de obter, armazenar e analisar dados.

Palavras-chave: Covid-19; Digitalização; Ciência de Dados; Inovação.

The digitalization of the rural and the democratization of data science: perspectives for application by agricultural producers

Abstract

The context of the Covid-19 pandemic, among its various consequences, accelerated the process of digitizing the field and stimulated the adoption of the internet by agricultural producers. The science of data (data science) with each passing day becomes more accessible and already begins to arrive in the most diverse organizations, including those related to agriculture and agribusiness. The paper sought to discuss the potential arising from the digitalization of the field and the use of data science with few resources by agricultural producers. It is a theoretical essay, based mainly on bibliographic research. The main results indicate that producers can benefit from numerous forms of digitalization and data science, however the main challenge is to increase the scope of internet access and the development of the culture of obtaining, storing, and analyzing data.

Key words*: Covid-19; Digitization; Data Science; Innovation.*

1 Introdução

A pandemia da Covid-19 atingiu o Brasil em um momento de recuperação da recessão de 2015/2016. As consequências da crise sanitária resultaram em um declínio expressivo da

[1]Doutorando em Administração no Programa de Pós-Graduação em Administração e Contabilidade (UFMS). https://orcid.org/0000-0001-7161-9693. Email: adm.leonardo.dresch@gmail.com
[2]Doutorado e Pós-doutorado em Economia Aplicada (UFV). Professor do Programa de Pós-Graduação em Administração e Contabilidade da UFMS. https://orcid.org/0000-0002-3677-1291. E-mail: adriano.figueiredo@ufms.br
[3]Doutorado em Economia Aplicada (UFV). Professora do Programa de Pós-Graduação de Administração e Contabilidade da UFMS. https://orcid.org/0000-0003-3961-2330. E-mail: bitencourtmayra@gmail.com

demanda externa, afetando também a demanda interna e restringindo a oferta. A proteção dos mais vulneráveis através da assistência social exigiu do governo federal um pacote fiscal, com prazo definido, estimado em R$ 713,4 bilhões, cujos efeitos devem cessar assim que o auxílio terminar. Tem-se um grande desafio à frente, relacionado ao alto desemprego, queda da renda e aumento da pobreza (BANCO MUNDIAL, 2020).

Schneider *et al* (2020), embasado em muitos documentos e relatórios, vislumbra para o Brasil um cenário bastante complicado. O aumento do desemprego, consequentemente diminuição da renda, e uma inflação acima da média para os produtos alimentícios desenham uma situação interna sensível. Entretanto, a análise aponta que a pandemia poderá ter efeitos benéficos e aumentar a oferta da produção e a inserção internacional do agronegócio do Brasil. São apontados os seguintes aspectos: taxa cambial favorável a exportações, com estimativa do dólar na casa dos R$ 5,00; possibilidade de acirramento na disputa comercial entre os Estados Unidos da América e China, que poderia ampliar as exportações brasileiras; e, por fim, a peste suína africana (PSA), que dizimou o rebanho suíno chinês gerando uma possível demanda de 13 milhões de toneladas de carne suína (todo o comércio internacional desse produto representa 9 milhões de toneladas) e poderá beneficiar o Brasil que possui potencial para suprir parte dessa necessidade de proteína animal.

Apesar do cenário de incertezas os produtores brasileiros, incluindo os integrados em cadeias agroindustriais e os que comercializam em circuitos curtos, encontram oportunidades na produção, distribuição e oferta. O aumento da eficiência é imperativo, e a digitalização e a democratização da ciência de dados podem representar uma das áreas para relevante melhoria.

O problema ao qual esse *paper* se propõe a discutir é: neste contexto de recente aceleração do processo de digitalização do campo proporcionado pela pandemia da Covid-19, de qual forma a integração digital e a ciência de dados de escassos recursos podem auxiliar os produtores?

O objetivo deste *paper* foi avaliar as possibilidades oriundas da digitalização do campo e da utilização da ciência de dados por produtores agropecuários. Especificamente, objetiva-se: i) discutir os exemplos de atividades proporcionadas pelo acesso a computadores com internet em três áreas: operação, planejamento, e, comercialização; e, ii) discutir as potencialidades criadas pelo acesso a técnicas da ciência de dados em outras quatro áreas: estatística convencional, algoritmos de classificação e agrupamento, análise de associação, e, série-histórica.

Contribui-se para a literatura da área ao indicar formas de integração do agronegócio no ambiente tecnológico 4.0. O trabalho desenvolvido é um ensaio acadêmico, fundamentado em

pesquisa bibliográfica, que busca relacionar o contexto da pandemia da covid-19 com a aceleração do processo de digitalização do campo e, aliado a popularização da ciência de dados (*Data Science*), como essa nova realidade pode ser apropriada pelos produtores agropecuários para aumento da eficiência dos seus negócios, mesmo que estes não possuam disposição ou recursos para investimento significativo na área.

Segundo Collis e Hussey (2005), a pesquisa pode ser classificada como: exploratória quanto ao objeto; quantitativa quanto ao processo; dedutivo quanto a sua lógica; e, básica quanto aos seus resultados. Busca-se informações sobre uma questão ou problema (exploratória), por meio do exame e da reflexão a partir das percepções do pesquisador para obter um entendimento de atividades sociais e humanas (qualitativa), em um problema de pesquisa com uma natureza pouco específica e buscando obtenção de entendimento sobre questões mais gerais (básica) e cujo estudo conduz a dedução de casos particulares a partir de inferências gerais (dedutiva).

O artigo está organizado em: i) introdução; ii) revisão da literatura (digitalização e a democratização da ciência de dados); iii) resultados e discussão; e, iv) considerações finais.

2 Digitalização e a democratização da Ciência de Dados

Segundo a Anatel (2021), 90,11% da população brasileira tem cobertura de banda larga móvel e 4.403 municípios possuem fibra óptica com uma meta de expandir para 4.883, dos 5.570 municípios brasileiros, até 2023. O aumento do acesso à internet faz parte dos Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU). O Programa Nacional de Banda Larga (PNBL), instituído pelo decreto n.º 7.175/2010 e o Programa Brasil Inteligente, criado através do decreto nº 8.776/2016 já possuíam em seus cernes a preocupação com as bases da universalização da internet no Brasil. A cobertura e a qualidade da internet no Brasil ainda devem ser positivamente impactadas com a chegada da tecnologia do 5G.

Sollitto e Venâncio (2020) afirmam que a pandemia acelerou o processo de digitalização do campo e coisas que demorariam anos estão acontecendo em meses. Para ilustrar essa progressão, citam a sétima edição da Pesquisa de Hábitos do Produtor Rural, realizada pela Associação Brasileira de Marketing Rural e Agronegócio (ABMRA), divulgada em 2017, com mais de 2,8 mil entrevistados, que apontava que 96% dos produtores tinham celulares, dos quais apenas 61% destes eram *smartphones* e apenas 42% afirmaram ter acesso à internet. Dos produtores com acesso, 53% dos entrevistados afirmaram nunca ter utilizado os meios digitais

para fazer qualquer cotação. Pesquisa da McKinsey, em 2020, apontou que 85% dos entrevistados no Brasil utilizavam o WhattsApp diariamente para resolver assuntos relacionados a produção, mesmo entre os produtores com menores níveis de alfabetização, e 71% usam canais digitais para questões ligadas à fazenda, apontando para um nível de digitalização entre os produtores rurais brasileiros maior que entre os americanos.

Costa (2020) conta o caso de produtores de orgânicos de Porto Alegre que foram surpreendidos com a chegada da Covid-19 em março de 2020. Decretos estaduais e municipais que restringiam a circulação de pessoas levaram a uma queda no movimento e nas vendas, gerando uma crise amenizada em parte pelas vendas conquistadas com o uso da tecnologia, como as redes sociais e os aplicativos de mensagens.

Os métodos científicos que alimentam a ciência de dados não são novos, como a estatística já presente em Pierre Simon Laplace (1749-1827) e Thomas Bayes (1701-1761), e a ciência da computação e o aprendizado de máquina, mais jovens, entretanto longe de serem novidades. A verdadeira mudança disruptiva está na evolução tecnológica na datificação, ou seja, na expansão acentuada da quantidade de dados agregados e digitalizados nas mais diversas áreas, bem como a democratização da análise de dados, que deixou de ser exclusividade de grandes empresas como Google, Yahoo, IBM ou SAS (IGUAL; SEGUÍ, 2017).

Amaral (2016) afirma que compreender a Ciência de Dados, ou mesmo o *Big Data*[4], parte da compreensão inicial sobre a sua matéria-prima, o dado, a informação e o conhecimento. Os dados são fatos coletados e normalmente armazenados, enquanto a informação é obtida a partir de dados analisados e com algum significado, já o conhecimento é a informação interpretada, entendida e aplicada para uma finalidade. A cada dia que passa os dados fazem mais parte da nossa rotina. Sites da internet rastreiam opções dos usuários, *smartphones* registram localização e velocidade, carros inteligentes registram padrão de condução, casas inteligentes coletam hábitos dos seus residentes (GRUS, 2016). Esses dados, e muitos outros, servem tanto ao setor privado quanto ao setor público e existe uma intensa corrida e disputa para compreendê-los, transformando-os em informações relevantes que podem ser aplicadas à tomada de decisão.

A recente revolução tecnológica nos trouxe uma imensa quantidade de dados e a necessidade de processar, armazenar, analisar e compreendê-los em grandes volumes e diferentes aplicações. Ganham força as áreas de estudo como a Inteligência Artificial (IA), o

[4] *Big Data*, expressão em inglês que pode ser traduzida como grandes dados. Segundo Cielen, Meysman e Ali (2016 p. 1, tradução nossa) "é um termo geral para qualquer coleção de conjuntos de dados tão grandes ou complexos que se torna difícil processá-los usando técnicas tradicionais de gerenciamento de dados".

Aprendizado de Máquina (*Machine Learning*) e a Ciência de Dados (*Data Science*). Os conceitos são frequentemente usados de forma intercambiável e até conflitantes entre si na mídia popular e na comunicação comercial, entretanto, apesar de relacionados, possuem diferenças (KOTU; DESHPANDE, 2019).

A inteligência artificial busca atribuir às máquinas a capacidade de imitar o comportamento humano, principalmente no que diz respeito às suas funções cognitivas. O aprendizado de máquina pode ser considerado um subcampo ou algumas das ferramentas da inteligência artificial, cuja finalidade é fornecer às máquinas a capacidade de aprender. A Ciência de Dados é a aplicação comercial do aprendizado de máquina, inteligência artificial e outros campos quantitativos, como estatística, visualização e matemática. É interdisciplinar e extrai valor dos dados, através da busca de estruturas úteis e significativas dentro de um conjunto de dados (KOTU; DESHPANDE, 2019). A figura 1 possibilita uma visualização a respeito dos conceitos e suas áreas de intersecção.

Amaral (2016) discute também a distinção entre o aprendizado de máquina (*machine learning*) e a mineração de dados (*data mining*). A primeira trata de algoritmos que buscam reconhecer padrões em dados, por outro lado, a mineração de dados refere-se à aplicação destes algoritmos em conjuntos de dados em busca de informação e conhecimento. Os padrões mais conhecidos para implementação de processos de mineração de dados são o CRISP-DM (*Cross Industry Standard Process for Data Mining*) e o KDD (*Knowledge-discovery in data-bases*). Ambos são semelhantes em suas etapas.

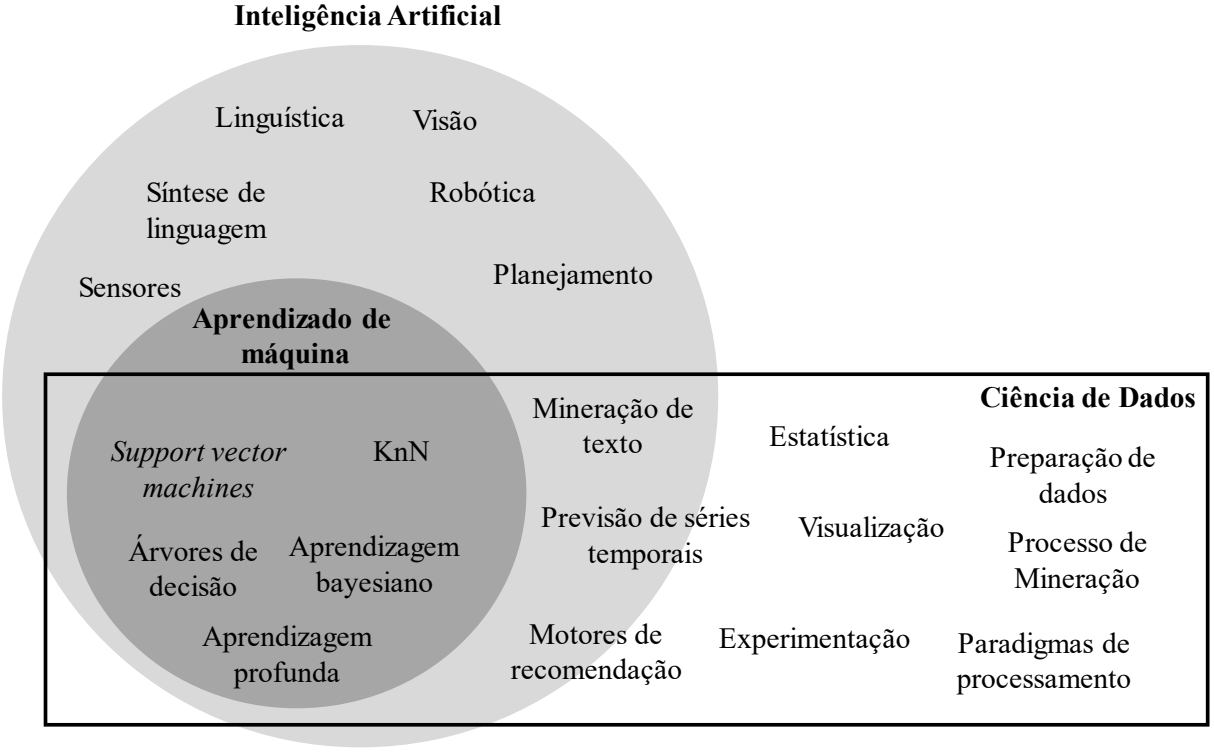
Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD tem como principal preocupação desenvolver métodos e técnicas para compreensão dos dados. A muitas vezes extensa base de dados é utilizada para se extrair automaticamente padrões que sejam devidamente compreensíveis e representativos.

> KDD, portanto, se caracteriza por ser um processo não trivial, que busca gerar conhecimento que seja novo e potencialmente útil para aumentar os ganhos, reduzir os custos ou melhorar o desempenho do negócio, através da procura e da identificação de padrões a partir de dados armazenados em bases muitas vezes dispersas e inexploradas. (THOMÉ, 2002, p. 11)

Os matemáticos e estatísticos definem o conjunto de ferramentas para modelagem e compreensão de um conjunto de dados complexos como aprendizagem estatística (*statistical learning*), subdividindo a área de estudo em dois tipos de técnicas, as supervisionadas e não supervisionadas. A aprendizagem estatística supervisionada busca prever ou estimar uma saída

com base em uma ou mais entradas, enquanto a não supervisionada possui entradas, mas nenhuma saída de supervisão e busca aprender e estruturar os relacionamentos a partir deste conjunto de dados. A aplicação é ampla e pode ser feita na área de negócios, medicina, astrofísica, política públicas e outras (JAMES *et al*, 2013).

Figura 1 – Conceito de Inteligência Artificial, Aprendizado de Máquina e Ciência de Dados



Fonte: KOTU e DESHPANDE, 2019 – tradução própria.

Cielen, Meysman e Ali (2016) complementam a conceituação de técnicas supervisionadas e não supervisionadas afirmando que, em termos gerais, são definidas pela quantidade de esforço humano necessário para coordená-las e como elas usam dados rotulados. É acrescida uma forma hibrida, uma técnica de aprendizagem semissupervisionada, a qual necessita de dados rotulados para encontrar padrões no conjunto de dados, mas ainda pode progredir no aprendizado mesmo se também houver dados não rotulados.

A análise de dados, portanto, consiste em um tipo de transformação dos dados em busca de conhecimento. As análises, a princípio, podem ser exploratórias, explícitas e implícitas. A análise exploratória utiliza de técnicas quantitativas e visuais para conhecer os dados antes de tentar analisá-los. As análises explícitas buscam informação e conhecimento disponíveis explicitamente nos dados através de tarefas normalmente de baixa complexidade, como a aplicação de um filtro ou ordenação de registros. A análise implícita, por outro lado, necessita de

alguma função mais sofisticada para se apresentar, como tarefas de aprendizado de máquina (*machine learning*) ou algum método estatístico (AMARAL, 2016). O autor dá um exemplo prático de análise explícita como uma empresa que busca saber quem são os seus funcionários que também atuam como fornecedores. A busca relacional entre as prováveis tabelas de dados de funcionários e fornecedores dará a resposta. Por outro lado, se a empresa quer prever quais novos clientes serão bons pagadores ou não para oferecer créditos especiais, as buscas nos históricos da empresa não demonstram relações facilmente compreensíveis, portanto, um algoritmo de classificação como o *Naïve Bayes*[5] possui a capacidade de prever essa informação com uma certa margem de erro.

O processo típico da ciência de dados consiste em seis etapas, a saber: i) estabelecer a meta da pesquisa, definindo o que, como e o porquê do projeto; ii) recuperação de dados brutos, internos ou externos; iii) preparação dos dados, detectando e corrigindo erros e transformando-os para utilização em modelos; iv) exploração de dados, buscando padrões, correlações e desvios com base em técnicas visuais e descritivas, fase que proporcionará os *insights* necessários que antecedem a modelagem; v) construção de modelo; e, vi) apresentação dos resultados e automatização da análise, se necessário (CIELEN; MEYSMAN; ALI, 2016).

As tarefas da ciência de dados, segundo Kotu e Deshpande (2018), são divididas em dez grupos: i) regressão; ii) agrupamento; iii) análise de associação; iv) detecção de anomalias; vi) motores de recomendação; vii) aprendizado profundo (*deep learning*); viii) previsão em séries temporais; ix) seleção de recurso; e, x) classificação. Amaral (2016) apresenta, no âmbito do aprendizado computacional, três principais tarefas, que são: i) classificação; ii) agrupamentos; e, iii) regras de associação (Quadro 1).

O cientista de dados (*data scientist*) deve ver os problemas da perspectiva dos dados. Precisa possuir habilidade analítica para o entendimento de problemas e negócios e envolve muito mais que apenas aplicação de algoritmos de mineração de dados (*data mining*), mas precisam possuir experiência e conhecimento na área em que estão investigando (WALLER; FAWCETT, 2013).

Amaral (2016) destaca os principais problemas que aparecem na construção de modelos de aprendizagem de máquina e cujo cientista de dados deve estar atento. Considerando um modelo de treino para determinação do modelo a ser testado com os dados de produção, os

[5] Em uma tradução comumente aceita, quer dizer “ingênuo” (do francês, *Naïve*) de Bayes. Em alguns idiomas tem-se o termo naive sem a trema.

principais problemas são: i) super ajuste do modelo (*overfitting*); ii) classe rara; iii) custo; iv) aprendizado baseado em instância; e, v) seleção de atributos.

Quadro 1 – Tarefas de aprendizado de máquina

Tarefas	Exemplos de tipos de algoritmos	Exemplo de Algoritmos
Classificação	Bayes	*Naïve Bayes* BaysNet
	Regras	*Party DecisionTable*
	Árvores de decisão	*Random Forest* J48
Agrupamentos	Por densidade	DBSCAN
	Baseado em protótipo	*K-Means K-medoids*
Regras de Associação	-	*Apriori FP Growth*

Fonte: Amaral (2016)

O super ajuste pode ser detectado quando o modelo em ambiente de desenvolvimento tem um desempenho excelente e quando testado em dados de produção sua performance cai muito, isso significa que o ajuste do modelo está adaptado demasiadamente aos dados de treino. A classe rara aparece quando os dados de treino não possuem a diversidade suficientemente representativa da realidade. O custo leva em consideração a finalidade ao qual o modelo foi construído, se para maximizar os verdadeiros positivos ou minimizar os falsos (e.g. seria mais importante, em um sistema de avaliação de crédito, conceder novos créditos a clientes ou reduzir o crédito dado aos maus pagadores). O aprendizado baseado em instância, característico de alguns algoritmos que a cada nova entrada recalculam seus resultados, requer maior capacidade de processamento e armazenamento. A última dimensão destacada pelo autor, referente a seleção de atributos, refere-se ao fato de que nem sempre adições de variáveis explicativas no modelo aumentam a sua performance, e, de fato, podem até atrapalhar. Alguns algoritmos lidam bem com isso, enquanto outros são afetados negativamente. Existem, entretanto, técnicas específicas que ajudam na seleção destes atributos.

2.1 Classificação

Os modelos de regressão assumem que a variável de resposta é quantitativa, porém, não raras são as situações em que precisamos que a variável de resposta seja qualitativa (categórica). O processo de prever variáveis qualitativas é conhecido como classificação. Prever a classificação de uma observação em um grupo delas envolve a previsão da probabilidade de que se enquadrem em uma das classes. As técnicas de classificação mais extensivamente utilizadas são: regressão logística, análise discriminante linear e K-*nearest neighbors* (vizinho mais próximo).

Entretanto, existem outras mais intensivas em computação, como *generalized additive models* (GAM), árvore de decisão (*decision tree*), *random forests*, *boosting* e *support vector machines* (SVM) (JAMES *et al*, 2013). A classificação é uma das tarefas desempenhadas em um contexto da ciência de dados que visa prever se o respectivo conjunto de dados pertence a uma classe predeterminada, sendo a previsão baseada nos aprendizados anteriores dos dados conhecidos do conjunto (KOTU; DESHPANDE, 2018).

Segundo James *et al* (2013), são alguns dos exemplos de problemas de classificação: i) uma pessoa chega ao pronto-socorro com sintomas (observações) que podem se enquadrar em três possíveis patologias (classes); ii) um serviço bancário *on-line* que possui alguns dados do seu cliente e do respectivo acesso (observações), e deve ser capaz de classificar a transação como lícita ou fraudulenta (classe). O classificador deve utilizar o conjunto de observações de treinamento $(x1, y1), \ldots (xn, yn)$ para construir um modelo de classificação que não seja adequado apenas para previsão da classe no conjunto de dados utilizados para treinamento, mas também em observações que não foram utilizadas para treinar.

Os classificadores Bayesianos, inspirados no Teorema de Bayes, são bastante populares e segundo Wu *et al* (2008) estão entre os dez tipos de algoritmos mais utilizados. São fáceis de construir, sem a necessidade de complicados esquemas de estimativas de parâmetros iterativos, e podem ser facilmente aplicados a grandes conjuntos de dados. Também são simples de interpretar, inclusive para usuários não muito qualificados e, finalmente, muitas vezes apresenta um desempenho superior. É conhecido como método ingênuo, por suposição de independência probabilística, apesar de uma vasta produção acadêmica já atribuir propostas de mudanças que afetam essas suas características de elegância, simplicidade e robustez.

O classificador *Naïve Bayes* é muito eficiente em termos de armazenamento e tempo de cálculo. A violação da suposição de independência tende a não prejudicar o desempenho da classificação para tarefas do mundo real. Os profissionais, entretanto, utilizam *Naïve Bayes* regularmente para classificação onde os valores reais das probabilidades não são relevantes, mas apenas os valores relativos nas diferentes classes. Outra característica que atribui vantagem a técnica é ser um natural aprendiz incremental, pois pode atualizar seu modelo sem reprocessar os exemplos antigos de treinamento quando novos dados surgirem (PROVOST; FAWCETT, 2016).

2.2 Análise de agrupamento (*Clustering*)

Segundo Hair *et al* (2009), pesquisadores frequentemente encontram situações que exijam um método objetivo para definir grupos de objetos homogêneos, sejam eles indivíduos, empresas, produtos ou mesmo comportamentos. O pesquisador procura uma estrutura natural entre as observações com base em um perfil multivariado. A técnica mais comumente usada para essa finalidade é a análise de agrupamentos. A análise de agrupamentos, também conhecida como análise de *clustering*, reúne indivíduos ou objetos em grupos tais que esses, no mesmo grupo, são mais parecidos uns com os outros do que com os objetos de outros grupos. A ideia é maximizar a homogeneidade dentro de grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos.

Segundo Kotu e Deshpande (2019) são exemplos de tarefas de agrupamentos: i) os clientes de uma empresa podem ser agrupados com base no comportamento da compra; ou ii) os eleitores em potencial podem ser agrupados em diferentes grupos para que os candidatos possam adaptar as mensagens para que ressoem melhor dentro de cada grupo. Uma diferença básica do agrupamento para a classificação é que se trata de aprendizado não supervisionado, não busca prever uma variável de classe de destino, mas simplesmente capturar os possíveis agrupamentos naturais nos dados. Todas as variáveis destacadas são agrupadas, independente da sua relevância ou pertinência, portanto, o utilizador do algoritmo deve ficar atento a possíveis distorções.

As diferentes técnicas de agrupamentos podem ser classificadas com base em como os dados se associam a um grupo identificado, assim o *cluster* pode ser: i) *Clusters* de particionamento exclusivo ou estrito, em que cada objeto de dados pertence a um *cluster* exclusivo; ii) *Clusters* sobrepostos, em que um objeto pode pertencer a mais de um *cluster*; iii) *Clusters* hierárquicos, em que cada *cluster* filho pode ser mesclado para formar um pai; e, iv) *Clusters* difusos ou probabilísticos, em que cada ponto pertence a todos os grupos de *clusters* com diferentes graus de associação.

As técnicas também podem ser classificadas conforme a sua respectiva abordagem algorítmica, considerando o tipo de relacionamento que eles utilizam entre os objetos de dados, que são: i) *Cluster* baseado em protótipo em que cada *cluster* é representado por um objeto de dados central, esse chamado de protótipo (e.g. centroide ou medoids); ii) *Cluster* de densidade, observado através da concentração de pontos de dados por unidade de espaço e separados por espaço esparso; iii) *Cluster* hierárquico, em que os *clusters* são criados em um processo ordenado

(de cima para baixo ou de baixo para cima) com base na distância entre os pontos de dados, dando origem a um dendograma ilustrativo desta hierarquia; e, iv) *Clustering* baseado em modelo, que se baseia nos modelos estatísticos e distribuição de probabilidade (e.g. Poisson ou Gaussiana) (KOTU; DESHPANDE, 2019).

Segundo Wu *et al* (2008) o algoritmo de agrupamento *K-means* é um dos mais utilizados nos modelos de Ciência de Dados. É oriundo de uma perspectiva diferente a do agrupamento hierárquico, que se concentra em similaridades de exemplos individuais e como estes se unem, focando nos próprios agrupamentos (os grupos), representados pelos centros dos seus agrupamentos (dado média, que são os centroides). O *k*, em *k-means*, é o número de agrupamentos que deseja encontrar nos dados.

A primeira etapa da implementação do algoritmo consiste na busca dos pontos mais próximos dos centros escolhidos (definidos aleatoriamente, com base no pré-processamento dos dados ou ainda determinados pelo usuário), resultando no primeiro conjunto de agrupamentos. A segunda etapa consiste em descobrir o centro real dos agrupamentos encontrados na primeira etapa. Os centros do agrupamento normalmente de deslocam nesse processo, sendo necessário recalcular quais pontos pertencem a cada grupos, repetindo o processo de cálculo dos centros do agrupamento de forma iterativa até que não existam mais mudanças (ou outro critério de parada seja atingido) (PROVOST; FAWCETT, 2016).

O algoritmo *k-means* é sensível à definição inicial dos centroides (não determinístico), não sendo garantia alguma de um bom agrupamento em uma única execução. Por esse motivo, normalmente são executadas várias vezes o algoritmo com diferentes centroides aleatoriamente definidos. Costuma ser um algoritmo eficiente e rápido, apesar das diversas iterações, mas tem alguns problemas. O desconhecimento do k a priori é um deles, mas existem formas de o analista buscar contornar esse problema (PROVOST; FAWCETT, 2016).

Entre as técnicas de agrupamentos *K-means*, Amaral (2016) traz ainda os algoritmos *Fuzzi C-Means* e o *K-medoids*. O primeiro é muito semelhante ao *K-means*, porém produz uma matriz com probabilidades de cada instância pertencer a um dos grupos, definindo a classificação absoluta com base na maior probabilidade. O *K-medoid*, ao invés de se basear em centroides, fundamenta-se em medoids, que são objetos, dentro do conjunto representativo de dados, que representam esse conjunto. Existem ainda muitos outros algoritmos de agrupamento, são apenas alguns dos exemplos o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), que diferente do *K-means* (que é uma técnica de *cluster* baseada em protótipo) é um

cluster de densidade, e o *self-organizing maps* (SOMs), uma técnica de *clustering* baseada em uma combinação de redes neurais e protótipo.

2.3 Análises de regras de associação

A análise de regras de associação (*Association rules analysis*), segundo Kotu e Deshpande (2019), mede a força da coocorrência entre um item e outro. Não busca prever a ocorrência de um item, como os algoritmos de classificação ou regressão, mas encontrar padrões utilizáveis nas coocorrências dos itens. É uma das áreas da aprendizagem não supervisionada que busca por padrões ocultos nos dados transformando-os na forma de regras reconhecíveis.

Segundo Amaral (2016), são frequentemente utilizados por sistemas de recomendação, incluindo grandes varejistas, sendo responsável por um terço das vendas da Amazon. Não existem grandes variações entre os algoritmos que produzem as regras de associação. Os mais populares são o *Apriori* e o *FP-Grow*.

Existem duas principais métricas utilizadas para avaliar a relevância das associações: suporte e a confiança. O suporte indica a proporção de instâncias que apresentam todos os itens, enquanto a confiança indica a proporção de instâncias que contendo um dos itens, também contém o outro. Um exemplo seria a compra do item A e do B, enquanto o suporte seria o percentual de compras que contenham A e B em relação ao total, a confiança seria o percentual de compras de A que contenham também o B, e vice-versa (AMARAL, 2016).

3 Os casos avaliados

Kalliandra (2021), uma empresa regional de tecnologia para o campo, ilustra em seu portfólio de serviços uma propriedade coberta por radiotransmissores e sensores. A sede da propriedade seria uma central de monitoramento agregando todas as operações. O pluviômetro instalado forneceria dados sobre a chuva em tempo real. Imagens de satélite seriam utilizadas para análise da área plantada, identificando precocemente possíveis anomalias geradas por falta de adubo, irrigação ou manifestação de pragas. As máquinas da fazenda seriam todas rastreadas sendo possível identificar através da central quais atividades estão sendo desenvolvidas. Diversos reservatórios de água teriam seus níveis monitorados por sensores, possibilitando uma maior racionalização desses recursos. Monitores de solo informariam possível déficit hídrico e possibilitariam avaliar se o pivô de irrigação está trabalhando de forma adequada. Os sensores

instalados no pivô de irrigação possibilitariam ainda o monitoramento da velocidade e da performance do equipamento, identificando qualquer problema. A estação meteorológica disponibilizaria dados climáticos regionais que aumentariam a capacidade de tomada de decisão para uma gestão racional da propriedade.

A empresa Premix (2021), cujo negócio é oferecer soluções em nutrição integradas à pecuária, lista diversas tecnologias aplicadas a essa atividade e que podem aumentar significativamente a eficiência das operações. A começar pelos *softwares* específicos para cuidar do rebanho e da propriedade, que proporcionam diversas informações com tempestividade, como: identificação individual dos animais; controle de pesagens; histórico de medicamentos; controle financeiro, estoque de medicamentos, insumos, suplementos e outros; registro de compra e venda da boiada; e, índices produtivos. É possível também, através de um *scanner*, medir o valor proteico-nutricional dos pastos para correção da dieta através de uma suplementação adequada.

A tecnologia NIRs pode ser utilizada para analisar a alimentação do rebanho. Ela utiliza a espectrometria e cálculos de regressão multivariada para obter os valores referentes aos níveis de proteína, umidade, gordura e fibra, entre outros parâmetros, em amostras de alimentos. Também é possível utilizar imagens de satélite para calcular com precisão o estoque de forragem superavitária ou deficitária dentro da propriedade, otimizando o uso de adubo para a correção do solo. A utilização de câmeras térmicas também proporciona a identificação precoce de animais enfermos e podem ajudar no manejo, através da identificação de locais com maior conforto térmico e até mesmo auxiliando no planejamento das estruturas dos piquetes de confinamento. É possível também utilizar câmeras para pesar o gado colocando-os em fila, durante a sua passagem em frente a câmera o programa faz a projeção das medidas do gado, cruzando informações em um banco de dados com milhares de referências, determinando o seu peso com uma considerável precisão (PREMIX, 2021).

Essas tecnologias ainda são apenas alguns exemplos. As possibilidades são ainda maiores incluindo sensores, drones, gps agrícola, tecnologias baseadas na mobilidade do 4g e aplicações de *big data*. Não são todos os produtores que possuem escala e recursos para a utilização desses meios, entretanto existem algumas atividades que podem ser desenvolvidas mesmo sem elevados custos. As discussões realizadas neste trabalho buscaram vislumbrar como o produtor, mesmo com poucos recursos tecnológicos disponíveis, já pode buscar benefícios oriundos desse novo paradigma produtivo baseado na tecnologia e na informação.

Os resultados da discussão realizada neste trabalho foram consolidados no quadro 1. O acesso ao computador com internet foi dividido em três grupos de exemplos de atividades: i) operação; ii) planejamento; e iii) comercialização. O acesso e domínio de ferramentas de Ciência de Dados foi organizado em quatro grupos de atividades: i) estatística convencional; ii) algoritmos de classificação e agrupamento; iii) análise de associação; e, iv) série-histórica.

Quadro 1 – Exemplos de atividades que podem ser realizadas por produtores agropecuários com avanço da digitalização e do acesso a Ciência de Dados.

Acesso ao computador com internet	Acesso a ferramentas de Ciência de Dados
➥ **Operação** - Atividades de apoio (administrativas) ➥ **Planejamento** - Previsão do tempo - Planejamento da produção - Pesquisa de insumos - Assistência técnica - Informações de mercado ➥ **Comercialização** - Aquisição de máquinas e insumos - Prospecção de clientes - Canais de comercialização	➥ **Estatística convencional** - Informações explícitas - Comparações de desempenho ➥ **Algoritmos de Classificação e Agrupamento** - Classificações e agrupamentos de clientes, fornecedores, áreas da propriedade entre outros; ➥ **Análise de associação** - Proposição de cestas de produtos ➥ **Série-histórica** - Identificação de sazonalidades e padrões temporais;

Fonte: Dados da pesquisa.

Os produtores agropecuários podem se beneficiar do acesso a computadores e a internet desempenhando atividades de apoio, especialmente as administrativas, como controle de estoque de insumos, lista de fornecedores e clientes e mesmo aquelas oriundas de exigências legais, como o controle de notas fiscais, folha de pagamento entre outras. O acesso as tecnologias

da informação e comunicação (TICs), durante a pandemia, foi fundamental para efetuar serviços bancários em tempos de restrições ao atendimento presencial.

O planejamento da produção também pode ser positivamente afetado pelo acesso as TICs. A previsão meteorológica, incluindo índices pluviométricos, por exemplo, são informações relevantes e que estão disponíveis sem custo e com níveis de acerto cada vez maiores. Ferramentas como o *Google Earth Engine* possuem potencial de avaliar grandes propriedades quanto a identificação de florestas, formações rochosas, nível dos lagos e rios, além da saúde da sua vegetação através da seleção de imagens de satélite de infravermelho já que "uma planta com mais clorofila refletirá mais energia do infravermelho próximo do que uma planta deficiente" (RODRIGUES, 2020, n.p). Com maiores recursos pode-se utilizar o *Normalized Difference Vegetation Index* (NVDI), ou Índice de Crescimento Vegetativo, cuja utilização por meio de imagens de satélites permitem a identificação de anomalias no desenvolvimento das plantas, que podem ser originadas por motivos diversos como problemas na irrigação, solo ou pragas.

O acesso a manuais de instrução, canais de dúvida de fornecedores de insumos e até mesmo serviços de assistência técnica estão muitas vezes disponíveis sem custo algum para o produtor. Também existe uma diversidade de canais que disponibilizam, organizam e compilam informações de mercado que podem apoiar o produtor em decisões sobre sua produção e comercialização. São alguns exemplos: o Cicarne (https://www.cicarne.com.br/), da Embrapa; o banco de dados da Companhia Nacional de Abastecimento (CONAB) (http://sisdep.conab.gov.br/precosiagroweb/); estudos do Instituto Brasileiro de Geografia e Estatística (IBGE) (https://sidra.ibge.gov.br/home/pms/brasil); Agência Nacional de Assistência Técnica e Extensão Rural (ANATER) (http://www.anater.org/); Sistemas do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) (http://sistemasweb.agricultura.gov.br/); estatísticas da *Food and Agriculture Organization* (FAO) (http://www.fao.org/statistics/en/); Estatísticas da *Organisation for European Economic Co-operation* (OCDE) (https://stats.oecd.org/); dados da *United States Departmente of Agriculture* (USDA) (https://usdabrazil.org.br/) além de muitos outros.

A grande variedade de possibilidades oriundas da digitalização na operação e planejamento da produção em propriedades rurais ainda são complementadas pelo acesso a novos canais de comercialização, como redes sociais, aplicativos de mensagens ou *home pages*, além da possibilidade de conseguir pesquisar e adquirir insumos para suas propriedades fora dos restritos e por vezes pouco concorrenciais mercados locais. São exemplos simples, usuais e sem

custo de como o acesso a um computador com internet podem aumentar a eficiência de operações em propriedades agropecuárias, mas de forma alguma essa breve discussão esgota as possibilidades.

A operacionalização da ciência de dados, por outro lado, exige um conjunto de conhecimentos específicos, porém amplamente disponíveis na própria internet. Colabora também uma variedade de *softwares* de código aberto. A estatística básica, com informações simples como a média e desvio-padrão, podem auxiliar na compreensão das coisas. Algumas amostras de duas áreas plantadas e tratadas com insumos e técnicas diferentes podem ser comparadas por meio de um diagrama de caixa (*boxplot*) ou ainda testados através de uma análise de variância para verificar qual é a mais eficiente. Uma percepção, que antes seria uma especulação, por meio da aplicação de conceitos e ferramentas da estatística pode se tornar clara para orientar as decisões e ações.

Algoritmos mais complexos da Ciência de Dados, incluindo aplicações de *deep learning*, já são amplamente utilizados em grandes operações industriais e agropecuárias cuja escala da operação suporta e se beneficia dos grandes investimentos. Máquinas autônomas já são uma realidade, envolvendo uma diversidade de sensores que indicam tanto aspectos da operação como do próprio funcionamento do equipamento.

Os produtores agropecuários de menor porte podem utilizar alguns destes recursos. Oliveira (2014), a título de exemplo, desenvolveu um *software* para efetuar reconhecimento de espécies florestais a partir de imagens digitais de madeiras utilizando uma aplicação de *deep learning* baseada em redes neurais convolucionais[6]. Existem também, amplamente disponíveis, aplicativos para reconhecimento de plantas, como o Google Lens, iNaturalist, PictureThis-Plant Identifier e PlantNet. Em breve a área da visão computacional deve se popularizar no campo, servindo para identificar pragas e deficiências nutricionais nas culturas, ou mesmo detectar necessidade de reposição de comida ou comportamento atípico de membro do rebanho. As possibilidades são muitas e a familiarização com as bases tecnológicas e seu barateamento estão em curso no momento.

Algoritmos de classificação e agrupamentos, já utilizados por organizações das mais diversas áreas para auxílio na tomada de decisão, podem ser facilmente modelados para organizar clientes, fornecedores, áreas da propriedade, tipos de rebanhos entre outros. A maior dificuldade está relacionada a cultura da coleta de informações e na capacitação do analista. A

[6] http://reconhecimentoflorestal.md.utfpr.edu.br/#/pt/classificador

análise de associação, também bastante popular em sistemas de indicações para compras on-line, podem ser utilizados para proposição de cestas de produtos.

As séries-temporais monitoram a mesma variável no tempo possibilitando a identificação de tendência, sazonalidade e aleatoriedade. A produção de uma propriedade ou seus custos podem ser as bases para o planejamento do fluxo de caixa que leve em consideração aspectos sazonais. Correlacionar variáveis, como por exemplo a produção e o índice pluviométrico, também possibilitam ampliar o conhecimento sobre as relações, e conforme as previsões disponíveis, orientar a uma tomada de decisão.

O domínio de técnicas da ciência de dados por parte do produtor rural e o desenvolvimento de uma cultura baseada em informações devem fazer parte da próxima etapa da modernização do campo. Sollitto e Venâncio (2020) destacam ainda que a digitalização do meio rural estimula a transição da liderança das propriedades para os mais jovens. Empresas como a Monsanto, multinacional voltada a agricultura e biotecnologia, e a Syngenta Digital, companhia Suíça controlada pela estatal chinesa ChemChina, já trabalham na área prometendo soluções tecnológicas que levam a um aumento de eficiência na utilização dos insumos com uma consequente redução de custos e aumento da produtividade.

4 Considerações finais

O problema ao qual esse ensaio se propôs a discutir foi a recente aceleração do processo de digitalização do campo, proporcionado pela pandemia da Covid-19, e de quais formas essa integração digital e a ciência de dados de poucos recursos poderiam auxiliar os produtores agropecuários.

Foram discutidos exemplos de atividades propiciadas pelo acesso a computadores com internet em três áreas: i) operação; ii) planejamento; e, iii) comercialização. Também foram discutidas as potencialidades criadas pelo acesso a técnicas da ciência de dados em outras quatro áreas: i) estatística convencional; ii) algoritmos de classificação e agrupamento; iii) análise de associação; e, iv) série-histórica. Mesmo sem um grande investimento em *softwares*, câmeras, radiotransmissores, sensores e outros equipamentos, é possível, por meio de um computador, *smartphone* e o acesso à internet, obter fotos de satélite, acesso a conteúdo técnico e de mercado, bem como aplicativos e programas gratuitos que podem resolver alguns problemas e auxiliar na tomada de decisão do produtor.

Em um país com dimensões continentais, o acesso universal a internet ainda não é uma realidade, porém é uma questão que avança rapidamente e os empreendimentos agropecuários devem estar preparados para essa nova realidade competitiva que já é vislumbrada por grandes empresas que atendem ao setor. Existe também uma já vista tendência de barateamento da tecnologia que hoje já atende a nichos mais capitalizados e que em breve estará disponível a maior parte dos produtores.

O trabalho, pela natureza exploratória, por obviedade não esgota o assunto, mas se propõe a lançar luz sobre esse novo paradigma produtivo ainda em processo de consolidação. Trabalhos abrangendo essa novíssima realidade ainda são poucos e este ensaio espera contribuir e estimular para que o fenômeno possa ser mais discutido. Para trabalhos futuros sugere-se a investigação de aspectos sociológicos envolvidos a resistência de alguns produtores em alterar o seu negócio, como os jovens estão participando desse processo e estudos de caso mais aplicados sobre como métodos de *deep learning* ou mesmo outras técnicas da ciência de dados podem auxiliar o produtor agropecuário.

Referências

AMARAL, F. **Introdução à ciência de dados:** mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

ANATEL. www.gov.br/anatel. **Overview of telecommunication in Brazil**, 2021. Disponivel em: <https://www.gov.br/anatel/pt-br/dados/acompanhamento/relatorios-de-acompanhamento/2021#R2021_3>. Acesso em: 8 Abril 2021.

BANCO MUNDIAL. www.worldbank.org. **Brasil:** aspectos gerais, 2020. Disponivel em: <https://www.worldbank.org/pt/country/brazil/overview>. Acesso em: 7 Janeiro 2021.

CIELEN, D.; MEYSMAN, A.; ALI, M. **Introducing data science:** big data, machine learning, and more, using Python tools. [S.l.]: Manning Publications Co, 2016.

COLLIS, J.; HUSSEY, R. **Pesquisa em administração:** um guia prático para alunos de graduação e pós-graduação. 2ª. ed. Porto Alegre: Bookman, 2005.

COSTA, F. D. www.ufrgs.br/jornal. **Pandemia acelera processo de digitalização de produtores orgânicos**, 2020. Disponivel em: <https://www.ufrgs.br/jornal/pandemia-acelera-processo-de-digitalizacao-de-produtores-organicos/>. Acesso em: 8 Abril 2021.

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

GRUS, J. **Data Science do zero:** Primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016.

HAIR, J. F. et al. **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009.

IGUAL, L.; SEGUÍ, S. **Introduction to Data:** A Python Approach to Concepts, Techniques and Applications. 1ª. ed. Cham: Springer, 2017.

JAMES, G. et al. **An introduction to statistical learning:** with Applications in R. New York: Springer , 2013.

KALLIANDRA. kalliandra.com.br/. **Manejo Inteligente para sua lavoura:** Tenha um manejo mais eficiente com acompanhamento de profissionais, ferramentas e sensores para uma decisão mais assertiva, 2021. Disponivel em: <https://kalliandra.com.br/>. Acesso em: 13 Abril 2021.

KOTU, V.; DESHPANDE, B. **Data science:** concepts and pratice. 2ª. ed. Cambridge: Morgan Kaufmann, 2019.

OLIVEIRA, W. D. **Software para reconhecimento de espécies florestais a partir de imagens digitais de madeiras utilizando deep learning**. Dissertação (Mestrado em Tecnologias Computacionais para o Agronegócio) – Universidade T ecnológica Federal do Paraná, Medianeira. 2018.

PREMIX. www.premix.com.br. **Tecnologia na pecuária:** como ter mais produtividade, 2021. Disponivel em: <https://www.premix.com.br/blog/tecnologia-na-pecuaria/>. Acesso em: 13 Abril 2021.

PROVOST, F.; FAWCETT, T. **Data Science para Negócios:** o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books, 2016.

RODRIGUES, R. B. br.granular.ag/. **O que é o Índice Vegetativo?**, 2020. Disponivel em: <https://br.granular.ag/blog/o-que-e-o-indice-vegetativo/>. Acesso em: 6 Abril 2021.

SCHNEIDER, S. et al. Os efeitos da pandemia da Covid-19 sobre o agronegócio e a alimentação. **Estud. av.**, São Paulo, v. 34, n. 100, p. 167-188, Dezembro 2020.

SOLLITTO, A.; VENÂNCIO, R. http://plantproject.com.br/. **Pandemia acelera a digitalização do campo:** processos que poderiam levar anos aconteceram em meses, 2020. Disponivel em: <http://plantproject.com.br/novo/2020/08/pandemia-acelera-digitalizacao-do-campo/>. Acesso em: 11 Abril 2021.

THOMÉ, A. C. G. http://docplayer.com.br/. **Redes Neurais:** uma ferramenta para KDD e Data Mining, 2002. Disponivel em: <http://docplayer.com.br/694203-Redes-neurais-uma-ferramenta-para-kdd-e-datamining.html>. Acesso em: 12 Abril 2021.

WU, X. et al. Top 10 algorithms in data mining. **Knowl inf Syst**, 2008. 1-37.