Combinando Inteligência Artificial e imagens de satélite para a previsão de sinistros agrícolas: Uma nota

Pedro Henrique Batista de Barros
Johns Hopkins University, EUA e Departamento de Economia, Universidade de São Paulo, Brasil

Adirson Maciel de Freitas Junior
Departamento de Economia, Universidade de São Paulo, Brasil

O setor agrícola está sujeito a adversidades provenientes de eventos climáticos, incidência de pragas, incêndios e variações de mercado, sendo, portanto, de suma importância a adoção de seguro rural para uma gestão adequada das atividades agrícolas. Entretanto, a existência de falhas de mercado inibe o desenvolvimento e a ampliação desse mercado, especialmente no Brasil. Nesse contexto, o principal objetivo desse artigo é propor uma metodologia inovadora que combina algoritmos de aprendizado de máquina com imagens de satélite ópticas e de radar para previsão de sinistros agrícolas que permita uma redução das assimetrias informacionais existentes no mercado brasileiro.

Palavras-chave. Inteligência Artificial, *Machine Learning*, Imagens de Satélite, Sensoriamento Remoto, Seguro Rural.

Classificação JEL. Q00, Q10, C81.

1. Introdução

O agronegócio tem um papel de suma importância no desenvolvimento econômico, tanto diretamente via aumentos de produtividade e renda, como indiretamente, ao induzir crescimento industrial e mudanças estruturais na economia. No Brasil, em especial, o agronegócio é estratégico para a economia do país ao contribuir de modo significativo para o superávit da balança comercial e ter uma participação importante no PIB, nas exportações e na oferta de trabalho (Tabosa et al., 2021).

Entretanto, o setor agropecuário está sujeito a diversos riscos que tornam a sua produção volátil e, em grande medida, incerta, destacando-se a possibilidade de problemas climáticos, pragas, incêndios e riscos derivados da dinâmica de mercado, que impactam nos preços dos insumos e no produto final (Mota et al., 2020; Tabosa et al., 2021; Tabosa e Vieira Filho, 2021). Além disso, o avanço das mudanças climáticas aumentará a frequência e a magnitude de eventos extremos e, portanto, os riscos do setor agropecuário (Vroege et al., 2021). De modo geral, as perdas do setor agropecuário afetam de modo adverso toda a sociedade, especialmente no que se refere a segurança alimentar, além de impactos generalizados na cadeia produtiva, nível de preços e nos

Pedro Henrique Batista de Barros pedrohbarros@usp.br
Adirson Maciel de Freitas Junior adirson@usp.br

 https://doi.org/10.5935/0034-7140.20230001

investimentos do setor (Silva et al., 2014). O risco é um componente inerente as atividades agrícolas, o que torna necessário, portanto, a adoção de medidas para sua administração e mitigação, com destaque para a contratação de seguro rural (de Oliveira Adami e Ozaki, 2012; Ozaki e Campos, 2017; dos Santos e da Silva, 2017; Mota et al., 2020; Vroege et al., 2021; Benami et al., 2021).

O seguro rural, ao reduzir as incertezas do produtor, permite uma maior previsibilidade da atividade, o que cria incentivos para a realização de investimentos que podem aumentar a produtividade e a produção agropecuária. Entretanto, há importantes falhas de mercado que inibem o desenvolvimento e ampliação do mercado de seguro rural (Tabosa et al., 2021). Uma das principais dificuldades é a existência de altos custos para o monitoramento e verificação do comportamento de um número elevado de produtores rurais, o que gera assimetrias de informações significativas e permite a ocorrência de moral hazard e seleção adversa (Ozaki e Campos, 2017; Tabosa et al., 2021; Benami et al., 2021). O Moral hazard se refere ao risco de o segurado adotar comportamentos que aumentem os riscos de perda, elevando o custo do segurador. A Seleção adversa reflete que o segurador, diante de assimetria de informação, estabelece o preço (prêmio) do seguro como uma função do risco médio, o que atrai apenas produtores com riscos mais elevados que a média (Benami et al., 2021; Vroege et al., 2021).

Na prática, os produtores transferem o risco para a segurador com o pagamento de um prêmio, fato representado pela apólice do seguro. Por isso, as seguradoras devem ser capazes de prever os sinistros agrícolas para assim administrarem adequadamente os riscos que lhe foram transferidos. Portanto, é de suma importância o desenvolvimento de métodos de monitoramento agrícola para mensurar a real dimensão das possíveis perdas, reduzindo assim a existência de assimetrias de informação. Nesse contexto, uso de geotecnologias, em especial o sensoriamento remoto, em conjunto com métodos de previsão, como o aprendizado de máquina, podem ser importantes instrumentos para uma melhor quantificação das possíveis perdas agrícolas e, consequentemente, dos sinistros (Ozaki e Campos, 2017; Mota et al., 2020; Klompenburg et al., 2020; Benami et al., 2021).

De fato, aplicações recentes de algoritmos de aprendizado de máquina em dados de sensoriamento, especialmente aqueles derivados de imagens de satélite, têm demonstrado um grande potencial. Isso se deve especialmente a capacidade desses algoritmos ajustarem uma função não linear a uma grande quantidade de dados e da ampliação das imagens de satélite com altas resoluções espaciais, espectrais e temporais disponibilizadas gratuitamente (Atzberger, 2013; De Leeuw et al., 2014; Klompenburg et al., 2020; Benami et al., 2021; Vroege et al., 2021). Nesse sentido, o objetivo geral do presente artigo é combinar algoritmos de aprendizado de máquina com imagens de satélite ópticas e de radar para a previsão de sinistros agrícolas. Apesar da combinação de sensores ópticos e de radar melhorarem de forma significativa a acurácia das previsões em agricultura, ela tem sido pouco adotada pela literatura (Meroni et al., 2021). Em seguida, aplicar a metodologia proposta para o mercado de seguro rural brasileiro para o ano de 2020, em particular, para o estado do Paraná, onde está localizado o maior número de apólices do país.

O seguro rural vem ampliando sua importância no Brasil nas últimas décadas, especialmente após a implementação do Programa de Subvenção ao Prêmio do Seguro Rural (PSR) em 2004, que passou a subsidiar parcela do prêmio pago pelo produtor, aumentando a oferta e a demanda por apólices no país (Silva et al., 2014; Mota et al., 2020). Dado o pequeno porte desse instrumento de política agrícola, quando comparado ao tamanho do mercado potencial, é possível afirmar com certa segurança que há um grande espaço para crescimento do seguro rural no Brasil. Ademais, o PSR tem uma pequena participação dentro dos gastos públicos federais destinados ao setor agropecuário, apesar de sua relevância dentro da política agrícola brasileira (dos Santos e da Silva, 2017).

Portanto, a principal contribuição do trabalho é desenvolver uma abordagem inovadora que pode servir como auxílio para o mercado de seguro rural do Brasil tanto diretamente no que se refere a previsão de sinistros agrícolas quanto indiretamente na determinação da taxa de prêmio a ser cobrada e no montante de recursos a manter como reserva para fazer frente as eventuais indenizações. Em outras palavras, propõe-se desenvolver uma metodologia para reduzir as possíveis falhas de mercado e incertezas existentes, permitindo, portanto, instrumentos para incentivar uma expansão sustentável do mercado de seguro rural no país.

Para se atingir os objetivos propostos, o artigo se encontra estruturado em cinco seções, incluindo esta introdução. A segunda caracteriza o mercado de seguro rural no Brasil enquanto a terceira seção detalha os materiais e métodos empregados. Os resultados e sua análise são realizados na quarta seção. Finalmente, na quinta têm-se as conclusões finais.

2. Mercado de Seguro Rural no Brasil

No Brasil, o crescimento do mercado de seguro rural ocorreu especialmente após a criação do Programa de Subvenção ao Prêmio ao Seguro Rural (PSR), com o Decreto n. 5.121/2004. Vale destacar que o programa se insere no arcabouço da “Lei Agrícola”, que estabelece que as atividades agropecuárias devem possuir uma rentabilidade adequada e que se deve minimizar as potências distorções e riscos. Dentre as principais diretrizes propostas pelo PSR, podemos destacar a universalização do acesso ao seguro rural, o incentivo a adoção de melhores práticas de gestão e garantir uma maior estabilidade financeira ao setor. Na prática, o programa subsidia uma parcela do prêmio pago pelo produtor no momento da contração do seguro rural e ao reduzir o valor do prêmio, cria-se incentivos para a aquisição da apólice, assim como aumenta sua oferta por parte das seguradoras, ampliando o mercado (de Oliveira Adami e Ozaki, 2012; Mota et al., 2020).

A concessão de subvenção possibilita uma maior previsibilidade e segurança para a atividade agropecuária, como também cria incentivos para a adoção de tecnologias redutoras de risco (dos Santos e da Silva, 2017). O PSR é considerado um dos pilares da política de administração de risco do MAPA, além de possuir um papel complementar nas políticas de crédito e comercialização. Na ausência de um mercado de seguro rural bem estabelecido, os produtores podem não conseguir honrar as suas dívidas de

crédito rural, o que pode gerar prejuízos significativos para o mercado financeiro e para o governo. Portanto, a ampliação do mercado de seguro rural também impacta positivamente o setor de crédito rural ao reduzir a probabilidade de calote ou renegociação de dívidas (Tabosa e Vieira Filho, 2021). Por isso, o seguro é um dos instrumentos prioritários da política agrícola atual (dos Santos e da Silva, 2017).

Percebe-se a ampliação da importância do programa no número de beneficiários, que ultrapassou os 100 mil em 2020, com montante assegurado de R$ 45 bilhões, uma subvenção de R$ 881 milhões e uma área de 3,67 milhões de hectares. Porém, a área segurada pelo PSR representou aproximadamente apenas 16,8% da área plantada do país enquanto o valor segurado atingiu somente 5,26% do faturamento da agropecuária nacional, No total, foram atendidos 62 tipos de culturas/atividades em 193.470 apólices com um percentual médio 30% de subvenção na taxa de prêmio. Houve um notório aumento nos recursos destinados ao PSR, que saltou de R$ 370,9 milhões em 2018 para R$ 881 milhões em 2020 (MAPA, 2020). Dentre as possibilidades para se ampliar o mercado de seguro rural no Brasil, destacam-se: (i) – estimular a concorrência no setor; (ii) – redução de assimetrias de informação; (iii) – aumento dos recursos do PSR e garantir estabilidade e previsibilidade nas concessões; (iv) – aperfeiçoar o ZARC (dos Santos e da Silva, 2017).

3. Material e Métodos

3.1 *Sensoriamento Remoto*

A crescente disponibilidade de imagens de satélite com alta resolução espacial, temporal e espectral, além de avanço em técnicas de baixo custo para monitoramento, têm possibilitado a sua aplicação cada vez maior na agricultura. O uso de Imagens de satélite permite o cálculo de informações biofísicas em curtos espaços de tempo e de forma barata, podendo ser utilizada para a realização de predições agrícolas (Atzberger, 2013; Weiss et al., 2020; Vroege et al., 2021). Entretanto, vale destacar que o uso do sensoriamento remoto na agricultura enfrenta dificuldades particulares muitas vezes não encontradas em outras atividades. Dentre elas, destacam-se padrões sazonais relacionados ao ciclo biológico das culturas, dependência das condições do terreno, climáticas e de manejo, fatores que podem introduzir grande heterogeneidade mesmo dentro de uma mesma cultura (Atzberger, 2013).

Muitas corretoras de seguro têm adotado o sensoriamento remoto para detectar condições adversas e prever eventuais perdas (Benami et al., 2021; Vroege et al., 2021). O grande potencial de utilização de dados de satélite no seguro rural pode ser resumido em três vertentes: (i) índices derivados de sensoriamento remoto possuem alta correlação com perdas reais; (2) baixo custo; (3) – permite a abertura de novos mercados ao reduzi assimetria de informações (De Leeuw et al., 2014).

O sensoriamento remoto permite analisar objetos na superfície terrestre sem o contato físico. Uma forma comum é o acoplamento desses sensores em plataformas orbitais, como os satélites. A escolha da resolução temporal, espacial e espectral dependem, em grande medida, do objetivo e do objeto de estudo. Vale destacar que,

quanto maior a resolução temporal, menor é a resolução espacial. Os sensores remotos utilizados são basicamente do tipo ativo ou passivo. Os passivos são normalmente de natureza óptica captando a refletância eletromagnética dos alvos. Os ativos, como os de radar e lidar, emitem sinais, com o sensor capando os efeitos de retroespalhamento. O presente artigo buscou combinar informações espectrais de imagens ópticas do Landsat-8 e valores de retroespalhamento de radar do Sentinel-1 para o ano de 2020.

Os dados de sensoriamento remoto são caracterizados por três tipos de resolução: (i) – espacial, tamanho do pixel; (ii) – temporal, frequência da imagem; e (iii) – espectral, número de bandas do espectro eletromagnético que é captado. Normalmente, o aumento de um tipo de resolução vem acompanhado com a diminuição de outra. Os dados obtidos a partir do sensoriamento remoto orbital contém informações úteis que podem ser utilizadas para estimar a quantidade de biomassa acima do solo, assim como a produtividade e a produção agrícola. Existe relações biofísicas importantes entre as informações espectrais derivadas de imagens de satélite e a produtividade das culturas agrícolas. Dentre essas relações, destacam-se as de biomassa que está relacionada com as bandas espectrais do vermelho que capta eficiência fotossintética e do infravermelho próximo que identifica o acúmulo de biomassa. Por isso, é comum a utilização de índices de vegetação que relacionam as informações espectrais, captadas pelos sensores, com a saúde, desenvolvimento e estágio fenológico da vegetação. Informações interpoladas de temperatura e precipitação também podem ser utilizadas para captar estresses durante o crescimento vegetativo.

No presente artigo, imagens do Landsat-8 e do Sentinel-1 foram selecionadas para compor a base dos dados para o treinamento do algoritmo. Inicialmente, para as imagens ópticas foi necessário mascarar as nuvens da área com base na banda de qualidade “pixel_qa”, cujo valor do pixel (322) não possui tais interferências. Feito isso, as imagens do Landsat-8 foram compostas pelos pixels medianos da coleção de imagens de 2020 na plataforma do Google Earth Engine (GEE). Vale destacar que as imagens utilizadas no GEE já possuem níveis de pré-processamentos realizados pelos repositórios proprietários das imagens. Nesse sentido, seguiu-se para a etapa de extração das variáveis que irão treinar o modelo algoritmo.

A partir das imagens ópticas do Lansat-8, o presente artigo se utilizou das seguintes bandas espectrais: (i) – azul; (ii) – verde; (iii) – vermelho; (iv) – infravermelho. O desenvolvimento e a saúde da vegetação podem ser obtidos a partir da combinação de informações de múltiplas bandas espectrais. Para treinar o algoritmo, utilizou-se os seguintes índices de vegetação: (i) – Difference Vegetation Index (DVI); (ii) – Ratio Vegetation Index (RVI); (iii) – Greenness Index (GI); (iv) – Normalized Difference Vegetation Index (EVI); (v) – Enhanced Vegetation Index (EVI); (vi) – Soil-Adjusted Vegetation Index (SAVI). Dentre elas, destacamos três: (i) – NDVI, combina informações da banda vermelha e do infravermelho para captar nível de biomassa. Apresenta saturação a partir de determinado nível; (ii) – EVI, combina o azul, vermelho e infravermelho para corrigir interferências atmosféricas e assim reduzir a saturação do NDVI; (iii) – GCVI, infravermelho próximo e verde para captar concentração de clorofila e, assim, identificar déficit nutricional.

O Sentinel-1 fornece observações da superfície terrestre numa frequência de 6 dias considerando o equador e em intervalos decrescentes conforme se direciona para os polos. O Sentinel-1 fornece coeficientes de retroespalhamento com polarização dupla na banda C ao nível de 10 metros de resolução espacial (Meroni et al., 2021). A partir das imagens de radar1 do Sentinel-1, tanto das órbitas ascendentes quando descendentes, foram utilizadas as seguintes informações captadas da sua banda C da polarização dupla resultante: (i) – co-polarização única VV (transmissão vertical/recepção vertical) e (ii) – polarização cruzada dual VH (transmissão vertical/recepção horizontal). Em seguida, também se calculou três indicadores, a partir dos valores SAR backscatterring: (i) – razão (VV/H); (ii) – diferença (VV-VH); (iii) diferença normalizada (NDI) – (VV-VH)/(VV+VH).

Os valores de retroespalhamento da vegetação são determinados, em grande medida, pelo ângulo e tamanho das folhas e pelo conteúdo de água. Para o solo, o valor reflete a sua umidade e rugosidade. Os indicadores, por sua vez, refletem quantidade de biomassa acima do solo e a estrutura tridimensional do complexo solo-dossel. Portanto, as observações do Synthetic Aperture Radar (SAR) do Sentinel-1 podem gerar informações complementares importantes, especialmente por permitirem identificar características alternativas do desenvolvimento das culturas e possibilitar a obtenção de dados mesmo com a presença de nuvens. Por isso, a combinação de ambos os sensores, o óptico e o de radar, melhora de forma significativa a acurácia dos algoritmos (Meroni et al., 2021; Vroege et al., 2021).

3.2 *Aprendizado de Máquina (Machine Learning)*

O Aprendizado de Máquina é um sub ramo da Inteligência Artificial e é utilizado principalmente por sua capacidade de identificar padrões, correlação e assim derivar informações úteis dos dados. De modo geral, o Aprendizado de Máquina tem se tornado um importante instrumento para a previsão agrícola (Klompenburg et al., 2020).

O objetivo do Aprendizado de Máquina (*Machine Learning*)[1] é construir modelos estatísticos para prever e/ou classificar algo de interesse. Na prática, existem dois tipos de algoritmos, os modelos supervisionados e não supervisionados. Neste artigo, propomos usar diferentes algoritmos baseados em modelos supervisionados para treinar uma inteligência artificial no objetivo de previsão de sinistros agrícolas. Em particular, para determinar o melhor algoritmo de aprendizagem de máquina para o objetivo proposto, testaram-se quatro alternativas complementares: *K-Nearest Neighbors (KNN)*, *Artificial Neural Networks (ANN)*, *Support Vector Machines (SVM)*, *Decision Trees (DT)* e *Random Forests (RF)*.

Para implementar essa abordagem, o primeiro passo é treinar os modelos para minimizar seu erro de classificação e evitar overfitting. Na prática, precisamos dividir nossa amostra de dados em dois: um para treinamento e outro para testes. Em outras palavras, este procedimento busca testar a validade do nosso modelo utilizando a amostra de teste para acessar o poder classificatório do modelo estimado. Para minimizar o viés

[1] Para detalhes sobre *Machine Learning* consulte Burger (2018). Para aplicações em sensoriamento remoto consulte Kamusoko (2019). Para aplicações agrícolas consulte Weiss et al. (2020).

potencial, é importante utilizar técnicas de amostragem para construir as amostras de treinamento e teste, as quais receberam 80% e 20% do total, respectivamente.

Para se certificar da robustez dos nossos resultados, nós utilizamos o método de validação cruzada k-fold, com especificamente 5-folds, para garantir que os dados de teste representem de fato nossa amostra de dados. Essa técnica divide a amostra de dados em k pedaços e cria, para cada pedaço, uma amostra de treinamento e teste e estima o modelo. Então, calcula-se a média dos erros previstos de cada um, em outras palavras, a validação cruzada permite acessar o grau de variação da classificação devido a amostragem. Além disso, vale destacar que adotamos o método recursivo de eliminação de variáveis (*recursive feature elimination*), que, na prática, é um algoritmo de otimização que tem como objetivo selecionar a melhor combinação de variáveis para se utilizar no treinamento do modelo.

Em seguida, após eliminado as variáveis não relevantes e o modelo devidamente treinado, é necessário comparar os resultados classificatórios do modelo treinado. Isso é realizado com a aplicação do modelo treinado na amostra de teste para, desse modo, obtermos uma visão geral sobre o desempenho das estimações numa amostra que não foi empregada no treinamento. Por fim, cada algoritmo de classificação foi comparado com base no nível de acurácia geral da classificação e no coeficiente de Kappa.

3.3 *Base de dados e análise descritiva*

A base de dados é composta pelas informações obtidas nos Relatórios Estatísticos do Seguro Rural do MAPA para o estado do Paraná no ano de 2020. Portanto, os dados se referem as indenizações ocorridas no âmbito do Programa de Subvenção ao Prêmio do Seguro Rural (OSR). No geral, a base de dados é composta de 9.548 contratos, totalizando 43 culturas e 14 seguradoras, sendo que aproximadamente 32% dos contratos apresentaram algum tipo de sinistro. No que se refere a abrangência geográfica, tem-se 354 municípios. A Tabela 1 mostra os principais municípios do Paraná em termos de importância segurada, além do número de apólices, prêmio total, subvenção e indenização, sendo notório, em especial, a participação da região Oeste do estado.

Tabela 1. Municípios do Paraná em termos de importância segurada em 2020.

Municípios	Apólices	Importância Segurada	Prêmio Total	Subvenção	Indenizações
Toledo	1600	R$ 213.842.329,36	R$ 15.635.603,35	R$ 5.221.301,08	R$ 7.862.927,35
Assis Chateaubriand	1410	R$ 206.514.307,69	R$ 15.186.123,21	R$ 5.037.692,21	R$ 4.884.220,06
Terra Roxa	681	R$ 153.892.889,35	R$ 11.142.322,95	R$ 3.535.314,64	R$ 5.001.014,54
Cascavel	1132	R$ 149.544.512,66	R$ 12.476.633,70	R$ 4.368.899,68	R$ 8.378.912,60
Mamboré	619	R$ 141.538.236,94	R$ 9.551.842,01	R$ 3.063.234,55	R$ 1.352.637,40
São Miguel do Iguaçu	632	R$ 124.530.217,92	R$ 8.522.328,69	R$ 2.811.511,84	R$ 537.214,79
Palotina	761	R$ 119.811.438,18	R$ 8.622.540,93	R$ 2.864.452,91	R$ 1.923.545,98
Luiziana	337	R$ 115.154.270,57	R$ 7.446.962,92	R$ 2.279.047,87	R$ 546.553,55
Londrina	844	R$ 110.232.003,95	R$ 9.160.888,73	R$ 3.149.275,62	R$ 635.639,53
Engenheiro Beltrão	683	R$ 102.978.165,53	R$ 6.736.108,17	R$ 2.222.158,17	R$ 2.432.839,56
Paraná	71132	R$ 10.477.746.635,76	R$ 747.811.141,96	R$ 243.805.219,22	R$ 288.487.792,57

A variável de interesse é a ocorrência ou não de sinistro, o que torna, portanto, a análise um problema de classificação. Vale destacar que a ocorrência de sinistros

é, em grande medida, correlacionada a produtividade e, portanto, também às bandas espectrais, aos valores de retroespalhamento e aos índices utilizados para treinar a inteligência artificial. Vale destacar que para melhor representar os potenciais riscos para a produção agrícola, considerou-se o ano safra para a composição das imagens de satélite, ou seja, utilizou-se o período de agosto de 2019 até julho de 2020. Em seguida, além de imagens com o valor mediano para todo o período, também foram utilizados os valores medianos considerando apenas do segundo semestre de 2019 e apenas o primeiro semestre de 2020 com o objetivo de captar diferenças entre as condições da primeira e da segunda safra. Também foi considerado os valores mínimos e máximos para o referido ano safra com a finalidade de captar amplitude de variação nos indicadores. Por fim, agregou-se para nível municipal os pixels considerando seu valor médio, procedimento necessário porque os contratos de seguro rural disponíveis Relatórios Estatísticos informam apenas o município, não detalhando sua localização geográfica. O procedimento resultou numa composição de 60 imagens que serão utilizadas para o treinamento do algoritmo.

Vale destacar que também foram considerados características municipais nas estimações, tais como: (i) a área média das propriedades rurais; (ii) – proporção de produtores com acesso ao mercado de crédito; (iii) – nível tecnológico do setor agropecuário; (iv) – proporção da população vivendo no meio rural; (v) – índice de Gini; (vi) – renda per capita; (vii) - abertura comercial; (viii) – escolaridade média do produtor rural; (ix) – área do município; (x) – densidade habitacional; (xi) – temperatura; (xii) – altitude; (xiii) – qualidade do solo; (xiv) – qualidade do solo; (xv) – precipitação; (xvi) – proporção de área agriculturável. O objetivo de incluir essas características é controlar pelas diferenças municipais e regionais existentes dentro do estado. No geral, essas informações foram obtidas no Censo Agropecuário de 2017 e Censo Demográfico de 2010 no Instituto Brasileiro de Geografia e Estatística (IBGE) e no Instituto de Pesquisa Econômica Aplica (IPEA).

Por fim, também se incluiu variáveis dicotômicas para cada seguradora e para as principais culturas seguradas, soja, milho e trigo, com o objetivo de controlar por efeitos diferenciados de exposição ao risco entre as seguradoras e culturas. De modo geral, é possível que cada seguradora adote uma política de gestão de risco diferenciada, gerando heterogeneidades na exposição. Por exemplo, é plausível que uma seguradora adote uma política conservadora na concessão de seguro rural enquanto outra seja menos cautelosa nesse quesito. Portanto, considerar essas diferenças é importante para a realização de uma adequada previsão dos sinistros agrícolas.

A Tabela 2 mostra as seguradoras em termos de importância segurada, além do número de apólices, prêmio total, subvenção e indenização. É possível verificar a presença de 14 seguradoras atuando no estado do Paraná e uma relativa pulverização na participação em suas respectivas participações de mercado.

As culturas agrícolas, por sua vez, possuem características fisiológicas e épocas de plantio diferentes, podendo, por exemplo, ser mais suscetíveis a pragas e eventos climáticos adversos que podem resultar em diferentes probabilidades de redução de produtividade. Além disso, o mercado de seguro rural pode se especializar na cobertura de algumas culturas agrícolas, em especial, aquelas com maior participação em termos

Tabela 2. Seguradoras no Paraná em termos de importância segurada em 2020.

Seguradoras	Apólices	Importância Segurada	Prêmio Total	Subvenção	Indenizações
Brasilseg	16405	R$ 2.770.722.786,76	R$ 200.786.496,60	R$ 67.658.642,83	R$ 44.873.672,14
Fairfax	11720	R$ 1.705.295.398,76	R$ 114.882.593,65	R$ 34.739.926,31	R$ 74.811.603,71
Newe	8889	R$ 1.219.137.112,30	R$ 87.414.650,74	R$ 28.511.855,82	R$ 24.466.877,17
Too	6801	R$ 1.082.694.958,94	R$ 72.396.838,23	R$ 22.281.667,20	R$ 29.591.309,75
Swiss Re	4069	R$ 806.107.402,07	R$ 55.865.844,87	R$ 18.306.721,62	R$ 8.589.388,40
Tokio Marine	5381	R$ 667.147.886,11	R$ 46.769.780,00	R$ 16.050.794,57	R$ 29.461.417,05
Allianz	3781	R$ 517.346.620,54	R$ 38.715.319,73	R$ 1.277.314,30	R$ 7.130.127,88
Sancor	3736	R$ 501.882.266,80	R$ 52.667.859,08	R$ 19.431.826,78	R$ 31.704.102,59
Mapfre	5665	R$ 545.137.784,74	R$ 25.215.889,17	R$ 7.308.488,46	R$ 4.586.539,86
Essor	2500	R$ 351.641.530,77	R$ 26.104.803,63	R$ 8.169.595,09	R$ 8.468.554,45
Aliança do Brasil	2808	R$ 153.810.668,10	R$ 12.103.162,26	R$ 41.416.783,06	R$ 8.548.209,82
Sompo	1558	R$ 105.757.164,72	R$ 7.378.346,54	R$ 2.262.103,04	R$ 3.536.572,53
Porto Seguro	826	R$ 77.376.494,55	R$ 4.591.688,00	R$ 1.273.692,50	R$ 10.159.711,26
Excelsior	405	R$ 64.988.560,60	R$ 2.914.869,46	R$ 889.807,64	R$ 559.708,96

de área plantada. A Tabela 3 mostra as 12 principais culturas em termos de importância segurada, além do número de apólices, prêmio total, subvenção e indenização. Dentre elas, destacam-se a soja com 38,28% dos contratos, seguida pelo milho com 35,44% e pelo trigo com 17,60%.

Tabela 3. Culturas seguradas no Paraná em termos de importância segurada em 2020.

Culturas	Apólices	Importância Segurada	Prêmio Total	Subvenção	Indenizações
Soja	38858	R$ 6.968.915.294,32	R$ 394.157.922,09	R$ 107.485.948,46	R$ 102.589.046,95
Milho 2ª safra	20417	R$ 2.174.675.046,96	R$ 229.962.780,84	R$ 90.994.194,35	R$ 13.400.076,00
Trigo	7821	R$ 651.386.326,68	R$ 85.836.389,98	R$ 33.483.673,87	R$ 28.722.934,92
Milho 1ª safra	1825	R$ 255.411.982,74	R$ 12.879.350,07	R$ 3.510.347,33	R$ 6.077.964,43
Feijão 1ª safra	728	R$ 78.596.208,38	R$ 6.084.969,97	R$ 1.486.642,02	R$ 6.788.751,04
Pecuário	119	R$ 66.326.765,50	R$ 1.126.518,85	R$ 425.722,82	R$ 448.566,26
Cevada	502	R$ 64.859.057,87	R$ 5.535.918,33	R$ 2.021.644,28	R$ 1.055.924,47
Floresta	38	R$ 54.444.281,18	R$ 771.158,12	R$ 232.946,64	R$ 128.804,32
Maçã	62	R$ 32.932.812,79	R$ 3.580.334,50	R$ 1.430.014,97	R$ 2.957.727,82
Batata	45	R$ 31.252.030,15	R$ 1.724.342,57	R$ 644.974,74	R$ 517.389,11
Café	143	R$ 19.568.501,82	R$ 1.027.774,10	R$ 265.575,98	-
Cana-de-açúcar	76	R$ 13.314.389,10	R$ 251.121,48	R$ 100.466,94	R$ 96.058,00

4. Resultados e Discussão

A presente seção traz os resultados encontrados e a sua discussão. Vale destacar que antes do treinamento da inteligência artificial propriamente dita foi necessário lidar com o fato das classes a serem classificadas serem desequilibradas. Considerando que a amostra final é composta de 9.548 observações, tem-se que a amostra de treinamento detém 6811 observações enquanto a de treino ficou com as restantes 1.702 observações. Entretanto, como já mencionado, apenas 32% dos contratos apresentaram algum tipo de sinistro, o que pode introduzir vieses no treinamento dos algoritmos devido ao desbalanceamento entre as classes. Para lidar com esse problema, utilizou-se o pacote ROSE no R, que utiliza uma técnica baseada em bootstrap para gerar amostras sintéticas balanceadas, o que resultou numa amostra final de treinamento com 4.676 observações com a mesma probabilidade de ocorrência entre as classes.

Para se obter os melhores resultados para a classificação das classes utilizados nesse artigo, cada algoritmo de aprendizado de máquina será comparado com base no seu nível de acurácia geral, coeficiente de Kappa e a área sobre a curva ROC, estimados na amostra de teste. Essas estatísticas são importantes para evidenciar a capacidade dos algoritmos de diferencias as classes, o que permite a seleção do modelo com melhor capacidade preditiva. Além disso, vale destacar que todos os algoritmos foram validados a partir do método de validação cruzada com 5-folds. Os resultados estão dispostos na Tabela 4.

Tabela 4. Resultados dos algoritmos de Aprendizado de Máquina.

	Acurácia Geral	Kappa
K-*Nearest Neihbor*	0,6693	0,3386
Artificial Neural Network	0,6405	0,2816
Support Vector Machine	0,6550	0,3108
Decision Tree	0,5848	0,1687
Random Forest	0,7135	0,4276

É possível verificar que o algoritmo com o melhor desempenho em tanto em termos de acurácia geral quanto de coeficiente de Kappa foi o *Random Forest* com 71,35% e 0.4276, respectivamente. Sendo assim, o resultado apresentou uma ligeira melhora de acurácia quando comparado ao trabalho de Mota et al. (2020) que atingiu uma acurácia geral média de 67,16% para seu melhor algoritmo. De qualquer modo, vale destacar alguns pontos e diferenças que podem explicar a pequena melhora de acurácia entre os trabalhos. Apesar da pequena melhora, é importante mencionar que o presente artigo considera apenas o estado do Paraná no ano de 2020 para o treinamento do algoritmo, ou seja, utiliza de modo limitado o potencial da base de dados de seguro rural que possui informações de 2006 a 2020 para todos os estados brasileiros que tiveram contratos. Dito de outra forma, espera-se que a acurácia dos algoritmos melhore com a ampliação da análise para todo o país e toda série temporal disponível, pois permitiria ao algoritmo explorar uma maior variação nos dados e assim apreender de forma mais efetiva a realizar previsões de sinistros agrícolas.

A Figura 1 mostra as 10 variáveis importantes em cada algoritmo estimado. Essa informação é importante para identificar quais foram as características que mais contribuíram para a realização da previsão dos sinistros agrícolas em cada modelo de aprendizado de máquina. Um ponto que merece destaque é que as variáveis mais importantes se alteram de modo significativo entre os algoritmos, indicando que diferentes modelos de aprendizado de máquina foram capazes de extrair e prever utilizando combinações alternativas de variáveis. Dito de outra forma, nenhum algoritmo conseguiu captar em conjunto toda a complexidade da relação entre as variáveis e os sinistros ocorridos e, por isso, o aprendizado não foi suficiente para se atingir uma acurácia mais elevada em previsões fora da amostra.

Nesse contexto, é possível conjecturar que, talvez, a adoção de métodos mais avançados de inteligência artificial, como o de aprendizado profundo (*Deep Learning*), possa

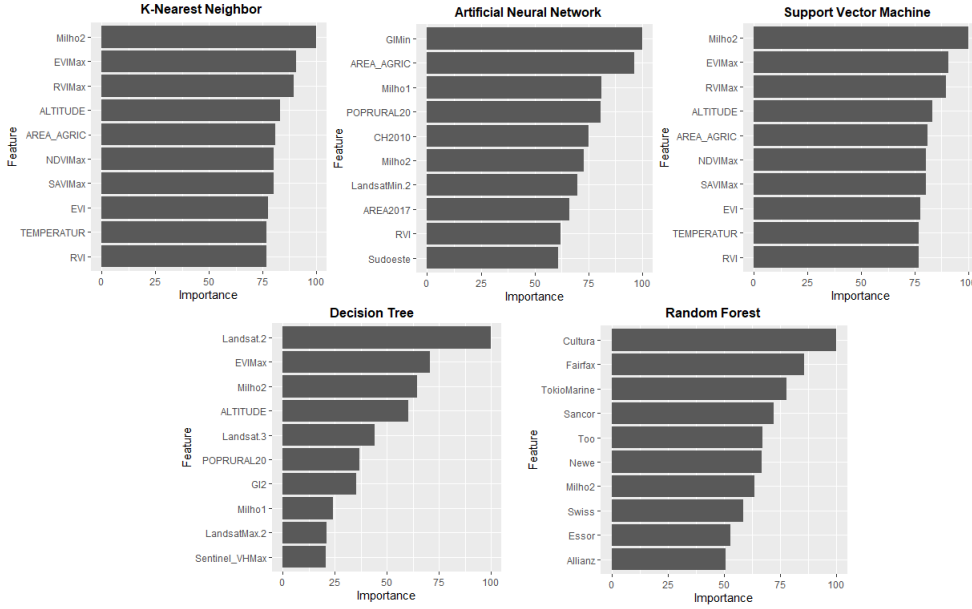

Figura 1. As 10 variáveis mais importantes em cada algoritmo.

ser uma forma de aprofundar o aprendizado do algoritmo, captando de forma mais completa as complexas variações existentes.

5. Conclusão

O presente trabalho propôs uma metodologia inovadora para a previsão de sinistros agrícolas combinando o uso de inteligência artificial e de imagens de satélite ópticos e de radar, juntamente a outras características importantes que possam afetar o risco agrícola. De modo geral, os resultados foram satisfatórios no sentido de evidenciar a factibilidade e a potencialidade do uso dessas tecnologias para a previsão de sinistro agrícola. De qualquer modo, os esforços aqui dispendidos podem se dizer preliminares, pois há diversos caminhos possíveis para a melhora dos resultados encontrados. A seguir, serão detalhados alguns caminhos possíveis para aprimorar os algoritmos e resultados encontrados.

A adoção de métodos alternativos de aprendizado de máquinas pode ser uma forma de melhorar a acurácia da previsão. Dentre eles, destacam-se tanto algoritmos recentes de Deep Learning, como Redes Neurais Convolucionais (*Convolutional Neural network*) e Redes Neurais Recorrentes (*Recurrent neural network*), quanto algoritmos clássicos de *Machine Learning*, como Análise de Discriminante (*Linear Discriminant analysis*), Regressões Ridge e Lasso, Splines de regressão adaptativa multivariada (*Multivariate adaptive regression spline*), Modelos lineares generalizados (*Generalizaed Additive Models*), classificadores Bagging e Boosting. Em suma, a metodologia adotada nesse artigo pode ser adotada com algoritmos alternativos de Inteligência Artificial com o objetivo de se buscar aquele com a melhor aprendizagem e acurácia de previsão.

Outro ponto importante que deve ser mencionado foi a utilização de uma base de dados limitada, quando comparado a disponibilidade de informações. O presente artigo treinou os algoritmos considerando o ano de 2020 para os sinistros agrícolas ocorridos no estado do Paraná. Portanto, um caminho natural para se melhorar os resultados aqui encontrados é a incorporação dos demais estados do país na análise, assim como a ampliação da base para toda a série temporal disponível. Na teoria, isso ampliaria a base de dados de treinamento, permitindo assim aos algoritmos captarem e aprenderem de forma mais completa as complexidades existentes entre as características utilizadas e os sinistros agrícolas. Em termos práticos, haveria um aumento na variação tanto em termos geográficos quanto temporais, o que possibilitaria controlar por fatores invariantes no tempo e/ou no espaço. Vale destacar que uma possível limitação dessa abordagem é o fato das imagens de radar do Sentinel-1 estarem disponíveis somente após 2014, ano de início do seu funcionamento. Portanto, apesar das informações de sinistro agrícola estarem disponíveis para o período de 2006 a 2020, há um trade-off em se adotar toda a série temporal devido a indisponibilidade de informações do Sentinel-1, fato que deve ser considerado.

Ainda no que se refere a base de dados, apesar dos algoritmos terem sido treinados a nível de contrato, a maioria das features utilizadas eram agregações em nível municipal devido a indisponibilidade de informações sobre a exata localização geográfica dos contratos. Em outras palavras, a base de dados utilizada somente disponibiliza informações sobre qual município o contrato foi realizado, não especificando sua localização detalhada. Portanto, um caminho para melhorar os resultados aqui encontrados seria a obtenção da localização detalhada dos contratos, o que permitiria calcular as informações derivadas das imagens de satélite especificamente para cada contrato. Esse procedimento certamente melhoraria a acurácia dos algoritmos utilizados tendo em vista que essa abordagem aumentaria a precisão e a variação dos dados utilizados.

A adoção de máscaras para diferenciar as culturas agrícolas dos outros usos e coberturas da terra, é outra possível abordagem para melhorar os resultados. O uso de máscara permite reduzir a análise para apenas os pixels que contém área agrícola, ou seja, permite a obtenção das informações das imagens de satélite somente para o subconjunto de área desejada, fato que tornaria os resultados mais precisos. De qualquer modo, vale destacar que no caso de utilização de série temporal para os dados de sinistro agrícola, o ideal é a utilização de máscaras anuais, pois é comum haver expansão/retração da área cultivada, assim como de rotação de culturas entre as safras.

Em suma, os resultados apresentados pelo presente trabalho demonstram a viabilidade e a contribuição da metodologia proposta para a previsão de sinistros agrícolas no Brasil. Dito de outra forma, a combinação de algoritmos de Inteligência Artificial e imagens de satélite se mostraram viáveis para utilização no mercado de seguro rural brasileiro. Inclusive, como detalhado nos parágrafos anteriores, há diversas possibilidades de melhora, tanto em termos metodológicos quanto tecnológicos, fato que abre a possibilidade de uma ampliação da aplicação e da importância dessa combinação no contexto do seguro rural.

Referências Bibliográficas

Atzberger, Clement (2013): "Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs," *Remote Sensing*, 5, 949–981. [2, 4]

Benami, Elinor, Zhenong Jin, Michael Carter, Aniruddha Ghosh, Robert Hijmans, Andrew Hobbs, Benson Kenduiywo, e David Lobell (2021): "Uniting remote sensing, crop modelling and economics for agricultural risk management," *Nature Reviews Earth & Environment*, 2, 1–20. [2, 4]

Burger, Scott V. (2018): *Introduction to machine learning with R*, Beijing, [China]: O'Reilly Media, 1st edition ed. [6]

De Leeuw, Jan, Anton Vrieling, Apurba Shee, Clement Atzberger, Kiros M. Hadgu, Chandrashekhar M. Biradar, Humphrey Keah, e Calum Turvey (2014): "The potential and uptake of remote sensing in insurance: a review," *Remote Sensing*, 6, 10888–10912. [2, 4]

de Oliveira Adami, Andréia Cristina e Vitor Augusto Ozaki (2012): "Modelagem estatística dos prêmios do seguro rural." *Revista de Política Agrícola*, 12, 60–75. [2, 3]

dos Santos, Gesmar Rosa e Fabiano Chaves da Silva (2017): "Dez anos do Programa de Subvenção ao Prêmio de Seguro Agrícola: proposta de índice técnico para análise do gasto público e ampliação do seguro," Discussion Papers 2290, Instituto de Pesquisa Econômica Aplicada - IPEA. [2, 3, 4]

Kamusoko, C. (2019): *Remote Sensing Image Classification in R.*, Springer Geography. [6]

Klompenburg, Thomas, Ayalew Kassahun, e Cagatay Catal (2020): "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, 177, 105709. [2, 6]

MAPA (2020): "Raio X do PSR - Relatório 2020. Ministério da Agricultura, Pecuária e Abastecimento (MAPA)," Acesso em 15/09/2021. [4]

Meroni, Michele, Raphaël d'Andrimont, Anton Vrieling, Dominique Fasbender, Guido Lemoine, Felix Rembold, Lorenzo Seguini, e Astrid Verhegghen (2021): "Comparing land surface phenology of major European crops as derived from SAR and multispectral data of Sentinel-1 and -2," *Remote Sensing of Environment*, 253, 112232. [2, 6]

Mota, Arthur Lula, Daniel Lima Miquelluti, e Vitor Augusto Ozaki (2020): "Predição de sinistros agrícolas: uma abordagem comparativa utilizando aprendizagem de máquina," *Economia Aplicada*, 24, 533–554. [1, 2, 3, 10]

Ozaki, Vitor e Rogerio Campos (2017): "Reduzindo a incerteza no mercado de seguros: uma abordagem via informações de sensoriamento remoto e atuária," *Revista Brasileira de Economia*, 71, 489–514. [2]

Silva, Alderir, Socorro Teixeira, e Vinicius Santos (2014): “Avaliação do programa de subvenção ao prêmio do seguro rural – 2005 a 2012,” *Revista de Política Agrícola*, 23, 105–118. [2, 3]

Tabosa, Francisco e José Eustáquio Vieira Filho (2021): “Análise espacial do programa de subvenção ao prêmio do seguro rural (PSR) e seu impacto na área cultivada e na produtividade agrícola no Brasil,” *Revista econômica do Nordeste*, 52, 27–43. [1, 4]

Tabosa, Francisco, José Eustáquio Vieira Filho, e Daniela Vasconcelos (2021): “Impacto do seguro agrícola na produtividade: uma avaliação regional no Brasil,” *Revista de Política Agrícola*, 30, 87–99. [1, 2]

Vroege, Willemijn, Anton Vrieling, e Robert Finger (2021): “Satellite support to insure farmers against extreme droughts,” *Nature Food*, 2, 215–217. [1, 2, 4, 6]

Weiss, M., F. Jacob, e G. Duveiller (2020): “Remote sensing for agricultural applications: A meta-review,” *Remote Sensing of Environment*, 236, 111402. [4, 6]