Utilizando Aprendizado de Máquina Explicável para Previsão de Severidade de Ferrugem Asiática da Soja

Christofer Daniel[1], **Guilherme Maturana**[1], **Eduardo H. M Pena**[1]

[1]Departamento de Computação – Universidade Tecnológica Federal do Paraná (UTFPR)
Campo Mourão, PR – Brazil

{christofer0888, guilhermear.maturana}@gmail.com, eduardopena@utfpr.edu.br

Abstract. *Asian rust is a disease caused by the fungus Phakopsora pachyrhizi; it has the soybean as one of its hosts, and its appearance depends on multiple enviromental variables. This paper proposes using state-of-the-art machine-learning models for predicting the severity of Asian rust on soybean crops. We combine geological and meteorological data with datasets on soybean crops to obtain a dataset that enables better exploring the capabilities of the models. In particular, we focus on explainable models. The extended dataset, in combination with the models, enabled achieving a good accuracy for prediction of Asian rust severity for several crops in several Brazilian cities.*

Resumo. *A ferrugem asiática é uma doença causada pelo fungo Phakopsora pachyrhizi; tendo a soja como um de seus hospedeiros, e seu aparecimento dependente de múltiplas variáveis do ambiente. Este trabalho propõe o uso de modelos de aprendizado de máquina estado-da-arte para prever a severidade da ferrugem asiática em lavouras de soja. Combinamos dados geológicos e meteorológicos com bases de dados sobre culturas de soja para obter um conjunto de dados que permita explorar melhor as capacidades dos modelos. Em particular, nos concentramos em modelos explicáveis. O conjunto de dados, em combinação com os modelos, permitiu atingir boa acurácia para previsão da severidade da ferrugem para várias plantações em várias cidades brasileiras.*

1. Introdução

A Soja é um grão com uma grande participação na economia mundial, podendo ser matéria prima para uma variedade de produtos, como ração de animais, óleos e alimentação humana [Bortolomedi 2022]. O Brasil, sendo um grande produtor rural, tem sua produção correspondendo a um terço da produção total de soja no mundo, com as exportações chegando a valores de 47 bilhões de dólares em 2022 [Piva 2022]. No entanto, existe um grande problema que aflige esse mercado, que é a ferrugem asiática. Essa doença, causada pelo fungo *Phakopsora pachyrhizi* provoca graves problemas de desenvolvimento para a planta, como a desfolha e a diminuição do tamanho do grão. Em casos

mais graves, a desfolha pode levar até a morte da planta, e causar perdas na produção de até 90% [Godoy 2022].

O problema da predição de ferrugem e a avaliação de sua consequente severidade é um tema que foi abordado previamente por outros autores, principalmente com o uso de modelos matemáticos [Del Ponte et al. 2006]. Modelos de simulação foram desenvolvidos para descrever processos biológicos que estimam o desenvolvimento da doença considerando disponibilidade de esporos, e que tentam prever a disseminação carregada pelo vento por longas distâncias[Zagui et al. 2022]. Foram também adaptados modelos matemáticos de doenças geral para a ferrugem asiática na soja e também modelos de regressão linear para analisar severidade baseado em temperaturas e umidade coletadas [Del Ponte et al. 2006].

Qualquer previsão sobre quão severa a doença se tornará durante uma safra pode auxiliar em tomadas de ações e decisões mais assertivas, pois o tratamento da ferrugem é feito com a utilização de diferentes agrotóxicos e não existe uma forma comprovada de encontrar o melhor momento de fazer a aplicação. Agricultores utilizam uma estratégia conhecida como método do calendário, em que em determinado estágio do desenvolvimento da planta, os produtores começam a fazer as aplicações [Barro and et. al. 2021]. Contudo, essa abordagem pode levar a um excesso de pulverizações, situação que faz os custos totais por safra ultrepassarem 2 bilhões de dólares, chegando a quase 3 bilhões, com uma média de 3,4 aplicações por safra [Embrapa 2023].

O objetivo deste trabalho é prever a severidade da ferrugem asiática.Tal previsão pode permitir medidas de combate à doença focadas nos pontos de risco, assim como análise de quais fatores levam a uma maior severidade da doença. Além disso, o sistema pode indicar quais os melhores agrotóxicos a serem utilizados em determinado momento, situação que colabora com o agricultor para cumprir protocolos de anti-resistência, que não recomendam a utilização de um mesmo agroquímico em mais de duas aplicações sequenciais [Godoy 2022].

Utilizamos modelos de aprendizagem de máquina capazes de prever a severidade da ferrugem asiática a partir de dados de cultivo, geológicos e meteorológicos. Esses modelos foram treinados em um conjunto de dados que mescla diferentes fontes de informações, com a principal sendo um estudo da Embrapa sobre a eficácia de diferentes agrotóxicos no combate da ferrugem asiática [Barro and et. al. 2021]. Esse conjunto passou por processos de fusão de dados com informações obtidas de outras fontes.

Assim, este artigo contribui no contexto da previsão da severidade da ferrugem asiática, propondo um modelo de aprendizagem de máquina explicável capaz de realizar a previsão da severidade em determinada localidade a partir de dados das plantações, geográficos e meteorológicos. A partir desses dados, e com uso nos modelos propostos, empresas e produtores podem implementar estratégias mais assertivas. Por exemplo, nossa solução permite identificar locais mais vulneráveis, o que pode ajudar no planejamento da utilização dos recursos disponíveis e diminuindo o custo total das produções.

2. Trabalhos Relacionados

O gráfico de favorabilidade climática da Fundação ABC utiliza coleta de estações climatológicas [Fundação ABC 2023]. Entretanto, essa ferramenta mostra a favorabilidade

climática para o desenvolvimento da doença em vez de uma estimativa para severidade. O trabalho [Zagui et al. 2022] utiliza técnicas de lógica *fuzzy* para estimar a vulnerabilidade do ambiente para o desenvolvimento da ferrugem. Diferentemente, nosso trabalho utiliza modelos de aprendizagem de máquina explicáveis que consigam prever a severidade da ferrugem. Além disso, buscamos prever a severidade da ferrugem e não a vulnerabilidade de um ambiente à infecção. O trabalho de [Cavanagh et al. 2021] utiliza um modelo de aprendizagem de máquina para a analisar como diferentes agrotóxicos interferem no desenvolvimento da ferrugem. O problema abordado nesse artigo é diferente do nosso, entretanto, ajuda mostrar como os pesquisadores estão engajados nos estudos relacionados à ferrugem asiática e como a inteligência artificial pode ajudar nesse processo.

3. Metodologia

Esse trabalho combina diversas fontes de dados em um formato tabular que permite explorar modelos de aprendizado de máquina eficientes para a previsão da severidade da ferrugem asiática. Os conjuntos de dados, integrados e tratados, e todo o código fonte dos experimentos podem ser acessados no link [1].

3.1. Coleta de Dados

Os dados de plantações utilizados para determinar a severidade da ferrugem asiática foram obtidos de testes cooperativos de fungicidas coordenados pelo Consórcio Antiferrugem [Barro and et. al. 2021]. Esses dados abrangem 10 estados brasileiros (Bahia [BA], Distrito Federal [DF], Tocantins [TO], Goiás [GO], Minas Gerais [MG], Mato Grosso do Sul [MS], Mato Grosso [MT], São Paulo [SP], Paraná [PR], e Rio Grande do Sul [RS]), com colheitas da safra de 2014/15 até 2019/20.

No estudo, a severidade é quantificada como uma média percentual para o lote de cultivo, levando em consideração a porcentagem da área afetada pela ferrugem nas folhas remanescentes das plantas, sendo atribuído o valor de 100% em casos de copas completamente desfolhadas [Barro and et. al. 2021]. Adicionalmente, são fornecidas informações detalhadas sobre os fungicidas utilizados, incluindo suas composições químicas e as quantidades aplicadas. Além disso, o estudo inclui lotes de controle nos quais nenhum fungicida foi utilizado para efeito comparativo.

Entre os dados utilizados para enriquecimento foram utilizados dados climáticos relacionados ao período da safra de cada cidade. Os dados estão disponíveis no pacote da linguagem *R*, *brclimr* [Saldanha et al. 2023]. Foram utilizados o desvio padrão e valores máximos, mínimos e médios mensais de precipitação, evapotranspiração, temperatura mínima, temperatura máxima, radiação solar, e umidade relativa do ar. Foram usados os valores climáticos referentes aos meses de novembro e dezembro do ano da safra, uma vez que esse é o período referente a pré-colheita.

Outras co-variáveis utilizadas para o enriquecimento foram as relacionadas ao clima da região. São elas os climas zonais, delimitações de regiões térmicas, e padrões de umidade e seca. Os dados foram coletados pelo Instituto Brasileiro de Geografia e Estatística e podem ser encontrados no portal de dados abertos do governo [IBGE 2023].

[1] `https://github.com/Fgarm/paper-soybean-rust-severity-prediction/`

3.2. Tratamento de Dados

Os dados passaram por processos de tratamento, para que apenas os dados necessários e relacionados às localidades desejadas fossem obtidos. Após isso, essas informações foram fundidas com o conjunto de dados principal da Embrapa de forma que fossem mantidas as relações de localidade e ano, assim, ao final desse processo, é obtido um conjunto de dados que relacionam solo, dados climáticos e agrotóxicos utilizados com a severidade da ferrugem asiática.

Dados referente às características climáticas de uma região foram obtidos nos repositórios do IBGE. Esses conjuntos são disponibilizados em arquivos no formato ".shp" que é um tipo de arquivo comumente utilizado por sistemas de informações geográficas. A informação contida dentro desse tipo de arquivo são descrições de como desenhar determinadas regiões e depois quais as informações relacionadas a determinada região.

Portanto, para obtermos os dados desejados foi necessário o desenvolvimento de um código capaz de usar as coordenadas de uma determinada região, informação encontrada no conjunto de dados da Embrapa, consultar cada uma das regiões descritas no arquivo do IBGE procurando em qual delas a localidade estava dentro, quando fosse encontrada a região, as informações eram retornadas. Esse processo foi executado para a obtenção dos dados referente aos aspectos climáticos de uma região e para os dados das características do solo, os dados obtidos eram então unidos com o conjunto de dados do Embrapa para ser relacionado com os dados de severidade.

Para a fusão dos dados obtidos do pacote *brclimr*, foi usado o código dos municípios (atribuidos pelo IBGE) onde as plantações ocorreram. Foram separado os valores de desvio padrão e valores máximos, mínimos e médios para cada diferente combinação de município, mês e ano. Foram selecionados os valores referentes a pré-colheita de cada safra, nos meses de novembro e dezembro, e os valores foram acrescentados ao conjunto de dados de severidade.

3.3. Modelos Utilizados

Durante esse estudo, foram realizados testes com diferentes modelos com a intenção de encontrar quais conseguiriam obter os melhores resultados. As características mais desejáveis dos modelos é a capacidade de encontrar comportamentos não lineares nos dados e de serem explicáveis, pois esse tipo de modelo oferece mecanismos para conseguirmos interpretar seus resultados e entendermos porque foi feita determinada previsão.

Um dos modelos utilizados é o *Random Forest*. Ele funciona a partir da combinação de várias estruturas chamadas árvores de decisão. Tais construções, sozinhas, geralmente não produzem boas previsões. Entretanto, a partir das estratégias de combinação de previsores empregadas no *Random Forest*, normalmente média ou voto, é possível obter resultados satisfatórios [Breiman 2001]. Essa situação se assemelha a um fenômeno conhecido como sabedoria das multidões.

Além disso, esse modelo tem grandes capacidades de explicabilidade. Com ele é possível obter informações como a importância de cada parâmetro e exportar uma representação visual de cada um dos *Weak Learners*. Isso possibilita visualizar os caminhos tomados pelo modelo para chegar em uma decisão. Finalmente, pode-se utilizar uma técnica conhecida como *SHAP* (*sampling-based approximation approach*) que permite observar o quanto um atributo dos dados está impactando o resultado final.

Outros modelos utilizados baseados em Árvores de Decisão são o *XGBoost* e o *CatBoost*. A diferença desses dois para o *Random Forest* é que eles utilizam técnicas mais robustas para a combinação dos resultados dos diferentes *Weak Learners*, chamada de descida de gradiente. Essa técnica usa cálculos matemáticos para reajustar os parâmetros dos *Weak Learner* para tentar melhorar o valor da previsão. Além disso, as capacidades de explicabilidade desses dois modelos são similares aos de *Random Forest*[Chen and Guestrin 2016][Dorogush et al. 2018]. Como exemplo, na Figura 1 é mostrada uma das Árvores de Decisão usadas internamente pelo modelo do *CatBoost*.

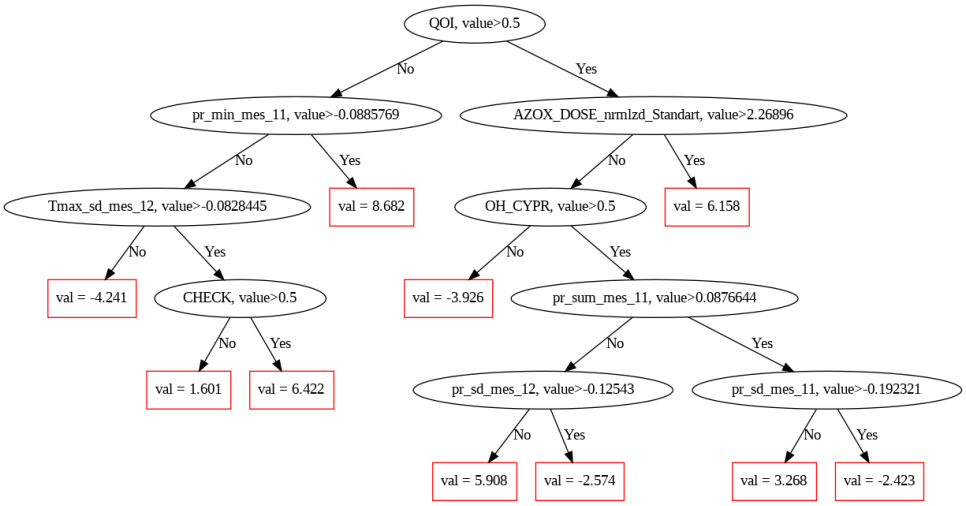

Figura 1. Visualização de um *Weak Learner* do modelo *CatBoost*

Outro modelo utilizado neste estudo foi a Rede Neural pois também apresenta boa capacidade de encontrar padrões não lineares em dados, entretanto seu funcionamento não é baseado em Árvores de decisão como os anteriores. As redes neurais foram criadas pensando em imitar o funcionamento do cérebro humano, ele é constituída de várias camadas de estruturas chamadas de neurônios que são conectados entre si por conexões sinápticas, e a maneira com que uma rede neural é treinada é calculando diferentes pesos para cada uma dessas conexões. O valor gerado no final é feito combinando os valores de cada conexão e seus respectivos pesos para cada camada [Wang et al. 2017].

As redes neurais não foram inicialmente construídas pensando em serem explicáveis, entretanto, por conta do interesse nesse tipo de características, mecanismos foram desenvolvidos para permitir essas capacidades. As formas de interpretação dos resultados de uma rede neural são a atribuição de importância, técnicas de visualização para entender a atividade dos neurônios nas diferentes camadas, além de ser possível a utilização do *SHAP*, apesar delas não terem suporte nativo como nos outros modelos.

3.4. Passo a passo do treinamento

No nosso *pipeline* de treinamento de modelos, adotamos um processo de pré-processamento de dados, incluindo a limpeza de duplicatas, a normalização de colunas e a eliminação de valores faltantes, com o objetivo de garantir a qualidade dos dados utilizados no treinamento. Além disso, o processo de pré-processamento é responsável pela combinação das diferentes fontes de dados para enriquecer e ampliar a diversidade das informações utilizadas no treinamento do modelo. Para otimização de hiperparâmetros,

utilizamos o método de *grid search*, que nos permitiu explorar diversas combinações de hiperparâmetros e selecionar aquela que resultou no melhor desempenho. Vale ressaltar que a métrica que utilizamos foi o Erro Quadrático Médio (RMSE), o qual nos forneceu uma medida da qualidade do nosso modelo em prever os valores de severidade.

3.5. Previsão da Severidade

Foram realizados testes para averiguar qual o modelo capaz de obter os menores valores de erro e buscar entender quais combinações de conjuntos de dados produzem os melhores resultados. Dessa forma, testamos combinações com diferentes fontes de dados. Além disso, realizamos testes para todo o conjunto e para cidades específicas, averiguando quais localidades estavam se adequando melhor aos dados.

4. Resultados

A partir dos testes realizados, foi possível realizar algumas análises. A primeira delas é comparar os gráficos das diferentes combinações para algumas cidades e ver como cada uma delas se comporta. Nas Figuras 2 e 3 podemos observar como variam os valores do RMSE para duas localidades diferentes, com uma delas obtendo uma melhora significativa, principalmente quando adicionando os dados do pacote *brclim* e outra com pouca melhora, mesmo com a combinação de todas as fontes de dados.

Nos gráficos a seguir, "T" representa os dados do pacote *brclimr*, "S" representam os dados de solo retirados do IBGE, "C" representa a base de dados de características climáticas retirados do IBGE, "A" representa a altitude da localidade da plantação e "B" é a base de dados sem acréscimos.

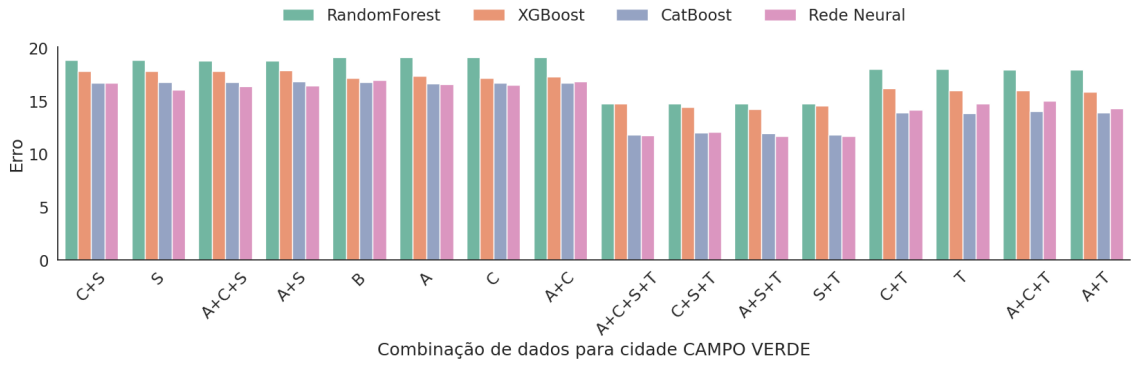

Figura 2. Comparação das diferentes combinações de dados para Campo Verde.

No eixo Y é possível observar valor do RMSE dos modelos estudados. Como no nosso trabalho estamos utilizando a técnica de regressão, esse valor é calculado obtendo a média das diferenças ao quadrado entre o valor previsto pelo modelo e o valor real, e quanto menor o valor, significa que o modelo está melhor ajustado e que as previsões estão próximas dos valores reais.

Outro resultado importante é sobre qual o desempenho dos modelos sobre todas as cidades. A Figura 4 traz essa informação.

Ao analisar o gráfico, percebemos que o acréscimo dos dados provenientes do pacote *brclimr* melhoram significativamente o erro de todos os modelos utilizados. Além

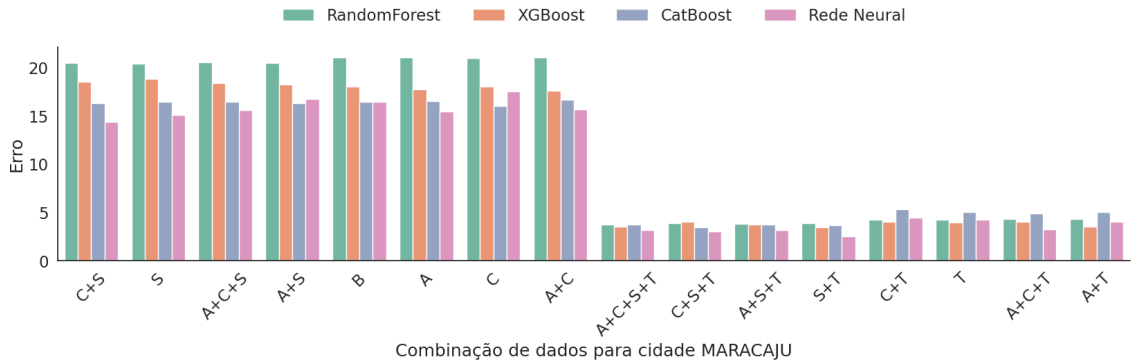

Figura 3. Comparação das diferentes combinações de dados para Maracaju.

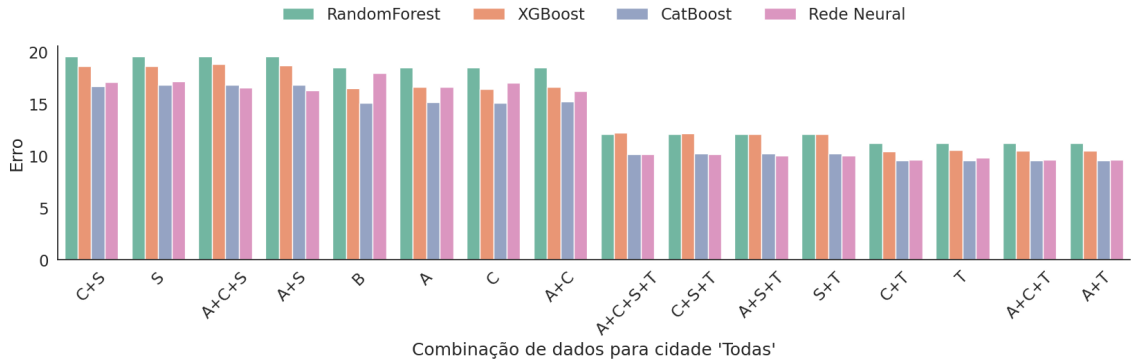

Figura 4. Comparação das diferentes combinações para todas as cidades

disso, é possível observar que em um dos casos, os dados do pacote *brclimr* sozinhos tiveram resultados melhores do que com todas as bases unidas, mostrando que alguns dados podem estar atrapalhando os modelos se ajustarem corretamente.

Outra situação relevante é que o modelo para todas as cidades não apresenta valores tão baixos como os mostrados em casos como o de Maracaju. Isso é por conta de um fenômeno causado pelos casos como o de Campo Verde, em que o modelo apresenta dificuldades para se ajustar a algumas partes do conjuntos de dados. Isso afeta a performance geral do modelo elevando o valor do RMSE. Tal fenômeno é detalhadamente estudado em [Chung et al. 2019].

5. Conclusões

Esse trabalho contribui com o contexto do desenvolvimento de ferramentas capazes de prever a severidade do desenvolvimento da ferrugem asiática. Soluções como essa auxiliam usuários (e.g., produtores, empresas) a escolherem estratégias mais assertivas quanto ao manejo da ferrugem em safras de soja. A partir dos resultados obtidos, foi possível observar como dados climáticos são importantes para que modelos de aprendizagem de máquina consigam boa capacidade preditiva para valores da severidade da ferrugem.

Agradecimentos. Este trabalho foi desenvolvido com suporte da Fundação Araucária.

Referências

Barro, J. P. and et. al. (2021). Performance of dual and triple fungicide premixes for managing soybean rust across years and regions in brazil: A meta-analysis. *Plant Pathology*, 70(8):1920–1935.

Bortolomedi, A. C. (2022). Qual a importância da soja? https://elevagro.com/conteudos/fotos/qual-a-importancia-da-soja.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Cavanagh, H., Mosbach, A., Scalliet, G., Lind, R., and Endres, R. G. (2021). Physics-informed deep learning characterizes morphodynamics of asian soybean rust disease. *Nature Communications*, 12(1):6424.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., and Whang, S. E. (2019). Automated data slicing for model validation:a big data - ai integration approach.

Del Ponte, E. M., Godoy, C. V., Canteri, M. G., Reis, E. M., and Yang, X. (2006). Models and applications for risk assessment and prediction of asian soybean rust epidemics. *Fitopatologia Brasileira*, 31(6):533–544.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *CoRR*, abs/1810.11363.

Embrapa (2023). Custo ferrugem, informações sobre a ocorrência da ferrugem-asiática e perdas pela doença nas safras de soja. Disponível em: `http://acacia.cnpso.embrapa.br:8080/cferrugem_files/ /764411951/Tabela_resumo_ferrugem_atual.pdf`.

Fundação ABC (2023). Monitoramento de doenças. https://sma.fundacaoabc.org/monitoramento/doencas_em_plantas/soja.

Godoy, C. V. (2022). Eficiência de fungicidas para o controle da ferrugem-asiática da soja, phakopsora pachyrhizi, na safra 2021/2022: resultados sumarizados dos ensaios cooperativos.

IBGE (2023). Instituto brasileiro de geografia e estatística - IBGE. https://dados.gov.br/dados/conjuntos-dados/cren_climadobrasil_5000/.

Piva, A. (2022). Valor da soja brasileira bate recorde em 2022. https://www.udop.com.br/noticia/2023/01/23/valor-da-soja-brasileira-bate-recorde-em-2022.html.

Saldanha, R., Akbarinia, R., Valduriez, P., Pedroso, M., Ribeiro, V., Cardoso, C., Pena, E., and Porto, F. (2023). *brclimr: Fetch Zonal Statistics of Weather Indicators for Brazilian Municipalities*. https://rfsaldanha.github.io/brclimr/.

Wang, H., Raj, B., and Xing, E. P. (2017). On the origin of deep learning. *CoRR*, abs/1702.07800.

Zagui, N. L. S., Krindges, A., Lotufo, A. D. P., and Minussi, C. R. (2022). Spatio-temporal modeling and simulation of asian soybean rust based on fuzzy system. *Sensors*, 22(2).