UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONEGÓCIO

FILIPE DA SILVA CÔRTES

Modelo de predição da ocorrência de ferrugem asiática na cultura da soja a partir de variáveis climáticas e clusterização.

Goiânia
2022

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese

2. Nome completo do autor

Filipe da Silva Côrtes

3. Título do trabalho

MODELO DE PREDIÇÃO DA OCORRÊNCIA DE FERRUGEM ASIÁTICA NA CULTURA DA SOJA A PARTIR DE VARIÁVEIS CLIMÁTICAS E CLUSTERIZAÇÃO

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **Klaus De Oliveira Abdala**, **Professor do Magistério Superior**, em 24/02/2022, às 16:06, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **FILIPE DA SILVA CORTES**, **Usuário Externo**, em 24/02/2022, às 17:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2707257** e o código CRC **1F979F1F**.

FILIPE DA SILVA CÔRTES

Modelo de predição da ocorrência de ferrugem asiática na cultura da soja a partir de variáveis climáticas e clusterização

Dissertação apresentada ao Programa de Pós-Graduação em Agronegócio da Escola de Agronomia da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Mestre em Agronegócio.

Área de concentração: Sustentabilidade e Competitividade dos Sistemas Agroindustriais.

Linha de Pesquisa: Meio Ambiente e Desenvolvimento Regional.

Orientador: Klaus de Oliveira Abdalla.

Goiânia

2022

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

CORTES, FILIPE DA SILVA
MODELO DE PREDIÇÃO DA OCORRÊNCIA DE FERRUGEM ASIÁTICA NA CULTURA DA SOJA A PARTIR DE VARIÁVEIS CLIMÁTICAS E CLUSTERIZAÇÃO [manuscrito] / FILIPE DA SILVA CORTES. - 2022.
LVI, 56 f.

Orientador: Prof. Dr. Klaus de Oliveira Abdalla.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-Graduação em Agronegócio, Goiânia, 2022.

1. Aprendizado de Máquina. 2. Modelos Preditivos. 3. Agricultura de Precisão. 4. Agrometeorologia. 5. Ferrugem Asiática da Soja. I. Abdalla, Klaus de Oliveira , orient. II. Título.

CDU 631/635

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **05/2022** da sessão de Defesa de Dissertação de **Filipe da Silva Côrtes**, que confere o título de Mestre em **Agronegócio**, na área de concentração em **Sustentabilidade e Competitividade dos Sistemas Agroindustriais.**

Aos vinte e quatro dias do mês de fevereiro de dois mil e vinte e dois, a partir das 14h, por meio de videoconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada "MODELO DE PREDIÇÃO DA OCORRÊNCIA DE FERRUGEM ASIÁTICA NA CULTURA DA SOJA A PARTIR DE VARIÁVEIS CLIMÁTICAS E CLUSTERIZAÇÃO". Os trabalhos foram instalados pelo Orientador, Professor Doutor Klaus de Oliveira Abdala (EA/UFG), com a participação dos demais membros da Banca Examinadora: Professor Doutor Alcido Elenor Wander (EMBRAPA), membro titular interno; e Professor Doutor Prof. Alexandre Bryan Heinemann (EMBRAPA), membro titular externo. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato aprovado pelos seus membros. Proclamados os resultados pelo Professor Doutor Klaus de Oliveira Abdala, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e quatro dias do mês de fevereiro de dois mil e vinte e dois.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Alcido Elenor Wander**, **Usuário Externo**, em 24/02/2022, às 15:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alexandre Bryan Heinemann**, **Usuário Externo**, em 24/02/2022, às 15:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Klaus De Oliveira Abdala**, **Professor do Magistério Superior**, em 24/02/2022, às 16:07, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2707256** e o código CRC **2FC55DFD**.

IMPACTO E RELEVÂNCIA DA PESQUISA

A ferrugem asiática é uma das principais doenças da soja e o controle dessa é feito principalmente com o uso de fungicidas. Porém, o uso desses fungicidas é feito de forma empírica, eles são aplicados de forma preventiva na plantação para que exista menos chance de a doença aparecer, pois os produtores rurais não possuem controle sobre o surgimento da ferrugem asiática na lavoura.

A pesquisa feita busca a identificação das condições climáticas propícias para o surgimento da ferrugem asiática. A implementação de tecnologia de agricultura de precisão é muito onerosa aos produtores, assim, na pesquisa buscou-se utilizar bases de dados públicas e abertas para a verificação da acurácia dos modelos preditivos, tornando acessível o acesso à tecnologia para esses produtores.

Na classificação das condições climáticas que proporcionariam a ocorrência de ferrugem asiática, a acurácia da classificação se mostrou muito positiva, sendo de 99%. Assim, a metodologia utilizada na pesquisa se mostrou certeira para a identificação das condições climáticas de ocorrência da ferrugem asiática na cultura da soja, utilizando-se os dados climáticos de região com pouca variabilidade climática.

Posto isso, a pesquisa poderá impactar a sociedade de forma significante, desde os produtores rurais até os consumidores finais. Isso porque a pesquisa gera a possibilidade de entregar uma ferramenta de apoio aos produtores rurais que poderá gerar maior sustentabilidade à produção agropecuária, reduzir o uso de fungicidas e definir melhores estratégias de controle da doença na cultura da soja, e ainda mitigar os impactos ambientais.

RESUMO

A ferrugem asiática é uma doença com impacto significativo na cultura de soja do Brasil. Apesar da grande relevância econômica da soja para o agronegócio brasileiro, são poucos os estudos sobre as condições que provocam a doença. Este trabalho objetivou a criação de um modelo preditivo considerando a influência de variáveis climáticas (temperatura, precipitação, umidade e radiação solar), embasados em um conjunto de dados de ocorrência da ferrugem, por meio da técnica de indução de árvores de decisão e regressão logística. O modelo foi criado com dados da ocorrência da doença nas cidades de Cristalina, Jataí e Rio Verde - GO nas safras de 2004/05 a 2016/17. Para cada registro de ocorrência (detecção), foi gerado um correspondente de "não ocorrência" sendo esse o trigésimo dia anterior ao dia da detecção, assumindo-se que nesta data não haveria presença de inóculo no campo. O conjunto de treinamento para a modelagem possui 10 variáveis totalizando 393 registros. O modelo preditivo foi proposto a partir da comparação de melhor desempenho entre os algoritmos de árvore de decisão e regressão logística. Após os resultados de acurácia obtidos (árvore de decisão 77,88%, contra 56,53% do algoritmo regressão logística), utilizamos o algoritmo de clusterização para agrupar os dados na fase de preparação de dados novamente comparamos o resultado entre árvore de decisão e regressão logística. Com o apoio da clusterização obtivemos a acurácia média na faixa entre 99 e 100% para árvore de decisão e 66,75 e 100% para regressão logística.

PALAVRAS-CHAVE: Aprendizado de Máquina. Modelos Preditivos. Agricultura de Precisão. Agrometeorologia. Ferrugem Asiática da Soja.

ABSTRACT

Asian rust is a disease with a significant impact on soybean in Brazil. Despite the great economic relevance of soybeans for Brazilian agribusiness, there are few studies on the conditions that cause the disease. This work aimed to create a predictive model considering the influence of climatic variables (temperature, precipitation, humidity and solar radiation), based on a dataset of rust occurrence, using the decision tree induction technique and logistic regression . The model was created with data on the occurrence of the disease in the cities of Cristalina, Jataí and Rio Verde - GO in the harvests from 2004/05 to 2016/17. For each occurrence record (detection), a corresponding "non-occurrence" was generated, this being the thirtieth day prior to the day of detection, assuming that on this date there would be no inoculum present in the field. The training set for the modeling has 10 variables totaling 393 records. The predictive model was proposed from the comparison of the best performance between the decision tree and logistic regression algorithms. After the accuracy results obtained (decision tree 77.88%, against 56.53% of the logistic regression algorithm), we used the clustering algorithm to group the data in the data preparation phase, again comparing the result between decision tree and logistic regression. With the support of clustering, we obtained the average accuracy in the range between 99 and 100% for decision tree and 66.75 and 100% for logistic regression.

KEYWORDS: Machine Learning. Predictive Models. Precision agriculture. Agrometeorology. Asian Soybean Rust.

LISTA DE FIGURAS

Figura 1. Conjunto de dados antes e depois da clusterização. 24
Figura 2. Processo KDD. 25
Figura 3. Folhas de soja infectada com o fungo Phakopsora pachyrhizi, causador da ferrugem asiática da soja, ilustrando as lesões resultantes da infecção. 26
Figura 4. Redução do erro médio quadrático conforme o número de clusters - "Teste do cotovelo". 33
Figura 5. Área plantada com soja (2004 – 2017) nos municípios analisados pela pesquisa. 37
Figura 6. Total geral de ocorrências por ano. 37
Figura 7. Matriz de correlação de ocorrência com variáveis climáticas - Captura de tela Jupyter Notebook. 37
Figura 8. Acurácia média em testes de regressão logística com e sem padronização escalar. 38
Figura 9. Gráfico de acurácia média em testes de árvore de decisão com e sem padronização escalar. 42
Figura 10. Gráfico de acurácia média em testes utilizando o target da clusterização. 46

LISTA DE TABELAS

Tabela 1. Atributos de cada ocorrência contida na tabela utilizada na pesquisa. 32
Tabela 2. Parâmetros alterados no modelo de regressão logística. 33
Tabela 3. Parâmetros alterados no modelo de árvore de decisão. 34
Tabela 4. Total de ocorrências por município (2006/2017). 36
Tabela 5. Medidas Estatísticas de Ocorrências (2006/2017). 36
Tabela 6. Medidas Estatísticas de Não Ocorrências (2006/2017). 37
Tabela 7. Resultados dos testes de regressão logística com k = 10. 40
Tabela 8. Resultados dos testes de regressão logística com k = 30. 41
Tabela 9. Resultados dos testes de regressão logística com k = 29. 41
Tabela 10. Métricas amostra k=30. 42
Tabela 11. Resultados dos testes de Árvore de Decisão k=10. 43
Tabela 12. Resultados dos testes de Árvore de Decisão k=30. 43
Tabela 13. Resultados dos testes de Árvore de Decisão k=29. 44

SUMÁRIO

1. INTRODUÇÃO 8
1.1 Objetivo Geral 15
1.1.1 Objetivos Específicos 15
2. JUSTIFICATIVA 15
3. REFERENCIAL TEÓRICO 16
3.1. Inteligência Artificial 16
3.2. Aprendizado de Máquina 17
3.3. Mineração de dados (Data Mining) 21
4. METODOLOGIA 27
4.1. Área de estudo 28
4.2. Procedimentos de Pesquisa 29
5. APRESENTAÇÃO E DISCUSSÃO RESULTADOS 34
5.1. Análise Exploratória 34
5.2. Regressão Logística 37
5.3. Árvore De Decisão 41
5.4. Clusterização K-means 44
CONCLUSÃO 45
REFERÊNCIAS 46

1. INTRODUÇÃO

O agronegócio brasileiro representa cerca de 26,6% do Produto Interno Bruto – PIB do país, (CEPEA, 2020), neste ambiente, se destaca a forte interação do setor agrícola com tecnologias avançadas da engenharia mecânica, química, genética, climatologia, o que permite considerá-lo um setor estratégico para alavancagem da economia no país.

Segundo Vale e Abdala (2018), as variáveis que compõem o ambiente agroclimático impõem um conjunto de restrições para a tomada de decisões do produtor, uma vez que constituem um ambiente que não pode ser controlado. Neste sentido, o uso de tecnologias, como a inteligência artificial, tem o potencial de ampliar a capacidade empreendedora de tomada de decisão neste ambiente, uma vez que, ao constituir ferramenta que permite modelagem do mesmo, internaliza as incertezas decorrentes do comportamento estocástico de tais variáveis, ampliando o índice de acerto em processos gerenciais.

A aplicação deste tipo de tecnologia no campo tem sido denominada como agricultura de precisão (AP) e, segundo Molin (2009, p.5), "trata-se de um conjunto de tecnologias aplicadas para permitir um sistema de gerenciamento que considere a variabilidade espacial da produção", assim, tem por objetivo otimizar a produção agrícola maximizando os resultados das colheitas.

A agricultura de precisão está relacionada com o uso de tecnologias como Internet das Coisas (IoT) e *big data* que tornam possível o monitoramento em tempo real permitindo análises dos fatores e variáveis que impactam nas lavouras, como o solo, clima, água, energia, e aplicação de insumos como sementes, fertilizantes, corretivos e os agroquímicos, tendo assim potencial de amenizar o impacto ambiental, uma vez que permite racionalizar a interação desses fatores.

Wolf e Wood (1997), destacam o desenvolvimento e a aplicação das Tecnologias da Informação e Comunicação na agricultura, para gerar ganhos econômicos e ambientais.Dessa forma, a AP (agricultura de precisão) constitui um caminho de busca pela eficiência e sustentabilidade nos empreendimentos agropecuários. O produtor que adota a AP como prática racionaliza ainda mais a utilização dos insumos ao promover a aplicação de forma correta, espaço temporal, e na quantidade certa, minimizando as perdas, o que permite elevar os benefícios econômicos e ambientais desta técnica.

A agricultura de precisão gera um volume de dados potencial para desenvolvimento

de sistemas de apoio à decisão. Segundo o Portal do Agronegócio (2018), o uso de inteligência artificial (IA) vai transformar o agronegócio, o Portal destaca que centrais de monitoramento estão cada vez mais presentes nas operações agrícolas. Essas centrais, utilizando inteligência artificial e Internet das Coisas, conseguem monitorar em tempo real toda operação. Dados do Ministério da Agricultura, Pecuária e Abastecimento (2017), apontam que 67% das propriedades agrícolas do país utilizam algum tipo de tecnologia, tanto para os processos de gestão dos negócios quanto para os processos de produção.

O Ministério da Agricultura, Pecuária e Abastecimento define a agricultura de precisão como "um sistema de gerenciamento agrícola baseado na variação espacial e temporal da unidade produtiva e visa ao aumento de retorno econômico, à sustentabilidade e à minimização do efeito ao ambiente" (BRASIL, 2009, p.31). Dessa forma, pode ser considerado um sistema de gestão para reduzir custos e aumentar a produtividade de modo sustentável com o mapeamento da produção.

De acordo com Molin, Amaral e Colaço (2014), a agricultura de precisão consiste em três estratégias de gestão, a primeira consiste em unidade de gestão diferenciada, ou seja, entender que cada parte da lavoura necessita de uma estratégia; a segunda tem por objetivo a redução de custos de insumos, baseada na adequação dos mesmos à variabilidade espacial; e a terceira é relacionada ao aumento da produtividade, conforme limites da lucratividade.

Dados do Ministério da Agricultura, Pecuária e Abastecimento (2017), apontam que 67% das propriedades agrícolas do país utilizam algum tipo de tecnologia, tanto para os processos de gestão dos negócios quanto para os processos de produção, o que permite a rápida adequação destas propriedades ao uso ampliado de tecnologias de informação com base em sensores. Dessa forma, o desenvolvimento de sistemas e modelos de apoio à decisão, utilizando-se de algoritmos de inteligência artificial, aplicados à predição e controle da produtividade, consiste em uma demanda do agronegócio.

As possibilidades e tendências do uso das tecnologias da informação e comunicação em agricultura de precisão são desenvolvidas juntamente com as tecnologias e métodos para armazenar e processar informações oriundas da combinação de hardware e software. O uso de ferramentas de análise permite identificar padrões para integração de toda cadeia produtiva agrícola, desde equipamentos/máquinas agrícolas à sistemas de gestão e apoio à decisão nos diversos níveis ao longo da cadeia.

Essas novas tecnologias possibilitam melhorias no processo produtivo, embora o monitoramento referente a pragas e doenças ainda seja realizado com envio de equipes especializadas ao campo, para coleta, contagem e identificação manual das pragas, o que torna o procedimento lento, e no caso de doenças, o processo de identificação também exige mais uma fase de laboratório ou incubação. Nesse sentido, as novas tecnologias de informação surgem para suprir essa demanda, com a possibilidade de garantir sustentabilidade, pelo fato de diminuir as aplicações de fungicidas minimizando o impacto ambiental, e gerando ganhos econômicos, ambientais e sociais.

O Sensoriamento Remoto aplicado à agricultura de precisão, segundo Shiratsuchi, Brandão e Vicente (2014), tem como princípio a combinação dos conhecimentos tradicionais e tecnologias de informação para uma produção inteligente. As técnicas de sensoriamento remoto fornecem informações através de imagens da lavoura de forma não-destrutiva, ou seja, sem existir contato físico com a plantação.

As imagens dos sensores podem ser obtidas através de satélites, aeronaves, veículos aéreos não tripulados, e máquinas agrícolas inteligentes, o que ressalta a importância da obtenção e tratamento dos dados coletados para constituição de diagnósticos como: estimativa da produtividade, detecção de pragas e doenças, previsão do tempo e avaliação hídrica das plantações.

A integração de veículos aéreos não tripulados (VANT's) na agricultura de precisão, contribui para redução do custo e otimização da produção, o tamanho e autonomia desses equipamentos são vantagens que surgem como uma importante opção para monitoramento de recursos naturais, meio ambiente, atmosfera, imageamento hiperespectral, observações de rios e lagos, bem como o imageamento de práticas agrícolas e uso do solo (Jorge, 2003).

Trabalhos recentes ressaltam, também, o potencial do uso de imagens de sensores orbitais e de veículos aéreos não tripulados (VANT's) no desenvolvimento de algoritmos de aprendizado de máquina (AM). Em sua pesquisa, Paiva (2021) obteve cerca de 93% de acerto para a classificação de culturas utilizando imagens de sensoriamento remoto e dados de EVI e NDVI, evidenciando o potencial de utilização da técnica e ferramental para a predição de produtividade e infestação de doenças.

Filho (2020) obteve cerca de 10% de erro ao utilizar dados de sensoriamento remoto captados por VANTs para identificação de ferrugem asiática na soja, especialmente em

plantas com 20 a 80% de severidade da doença. Para a predição de produtividade na sojicultura, Richetti (2018) utilizou uma série histórica de imagens de sensoriamento remoto e realizou testes com quatro algoritmos distintos, seu melhor resultado foi 3,52 kg ha^{-1} de erro médio acima do valor real de produtividade da soja no estado do Paraná.

O trabalho desenvolvido por Schwalbert (2019) trouxe resultados que reforçam o potencial do uso de modelos de predição como uma ferramenta de apoio à decisão. Estes modelos preditivos, baseados em imagens de satélite e variáveis meteorológicas, anteciparam a produtividade da cultura da soja em até 70 dias com um erro médio absoluto de 0,42 Mg/ha. Temperatura do ar, da superfície do dossel e déficit de pressão de vapor são variáveis que melhoraram o desempenho dos modelos em relação aos modelos baseados em apenas índices de vegetação - NDVI e EVI.

Os modelos de previsão de produtividade desenvolvidos por Schwalbert (2019), apresentaram baixa capacidade de generalização fora dos limites onde foram ajustados, uma vez que os parâmetros foram delimitados em nível de área agrícola. Apesar desta desvantagem, a especificidade dos modelos ainda auxilia na gestão e manejo da cultura, nos locais de concepção dos mesmos, com melhores insights que um modelo generalista poderia oferecer.

Os resultados das pesquisas de revisão sobre o tema indicam que o uso de modelos computacionais são ferramentas que podem auxiliar a tomada de decisão dos produtores referente à ações relacionadas à prevenção à doenças.

O controle de doenças é um dos maiores problemas existentes na agricultura, doenças como a ferrugem asiática (*Phakopsora pachyrhizi*) são de difícil contenção, propagam-se rapidamente nas lavouras podendo acarretar perdas em até 90% da produção. Durante as safras de 2001/02 a ocorrência da ferrugem asiática foi constatada nos Estados do Rio Grande do Sul, Paraná, São Paulo, Mato Grosso, Mato Grosso do Sul e Goiás, com perdas de até 70% da produtividade nas lavouras atingidas (YORINORI e PAIVA, 2002), sendo, atualmente, uma das principais doenças na cultura de soja, com custo aproximado por safra de US$ 2 bilhões. (CEPEA, 2019)

Em 2003, a epidemia de Ferrugem da Soja foi observada em 90% dos campos no país, com perdas estimadas em regiões como Mato Grosso e Bahia de US$ 487,3 milhões, e os custos com os fungicidas para duas pulverizações foram projetados em cerca de US$ 544 milhões (YORINORI, 2005). Os danos causados pela doença sobre a produção são

observados no menor peso das sementes, redução dos números de vagens e sementes e, em casos severos, queda das vagens.

Segundo registros do Portal Mais Soja (2018), na safra 2017/18, o país já registrava aumento de 20% de ocorrência de ferrugem asiática, a doença já havia sido detectada em todos os estados produtores. A elevada presença da ferrugem nestas safras, tem exigido que o produtor considere o uso de diversas ferramentas no combate à doença, objetivando garantir os níveis de produtividade e redução de custos, inclusive os ambientais, decorrentes da aplicação de agrotóxicos. Uma das alternativas são os cultivares tolerantes à ferrugem asiática, que consistem em uma tendência entre os grandes produtores. Esses novos cultivares estão disponíveis no mercado desde 2008 e auxiliam no controle da ferrugem asiática da soja; apesar de conseguirem retardar o avanço da doença, os cultivares ainda são de baixa disponibilidade no mercado e não dispensam o controle químico.

O combate à doença, devido os severos danos causados, associado ao fato da baixa disponibilidade no mercado de cultivar resistente a Ferrugem, é realizado de forma preventiva e o controle químico se torna o método mais eficaz e imediato, seguindo um calendário criterioso para pulverização dos fungicidas, que tem custos significativos para os produtores, além disso, o patógeno *Phakopsora pachyrhizi* tem adquirido continuamente resistência à determinados fungicidas (Reis et al, 2020). Diante desse cenário o Ministério da Agricultura Pecuária e Abastecimento (MAPA), passou a recomendar a utilização de misturas dos fungicidas para o controle da ferrugem (GODOY et al., 2012).

O controle químico desta fitopatologia gera um conjunto de custos, que englobam desde os custos mecânicos da pulverização (custo máquina), além do fungicida e do adjuvante para os fungicidas em que é recomendado seu uso, até os ambientais e humanos devido aos impactos na saúde destes organismos. Os custos na região Centro-Sul referente a safra de 2017/2018 com defensivos foram de R$ 716,20 por hectare para soja convencional, representando 41,28% do total das despesas de custo da lavoura. E para soja transgênica o custo por hectare foi de R$ 660,77, o que representou 34,06% de despesas de custeio da lavoura (IMEA, 2016).

Outra estratégia de controle que vem sendo adotada pelos produtores é o Manejo Integrado da Cultura, que exige treinamento para identificação precoce da doença, uma vez que em condições climáticas favoráveis a presença do fungo é contínua, exigindo constante monitoramento devido às variações climáticas e diferentes zonas de cultivo no país, o que

dificulta uma recomendação genérica de controle que atenda todas as regiões. O Manejo Integrado da Cultura consiste em levantamentos que abrangem desde o tamanho da área a ser semeada até a aplicação criteriosa do fungicida, em uma sequência de etapas, conforme ESALQ (2006).

Devido ao enorme prejuízo causado aos produtores pela ferrugem-asiática, uma das iniciativas para o combate da doença foi a criação do Consórcio Antiferrugem, uma parceria de ações público-privada, onde o produtor pode obter informações atualizadas sobre a doença, mapa da dispersão, informativos de risco e laboratórios credenciados. Tudo pode ser acessado através de plataforma web e aplicativos móveis de modo simples e intuitivo; as informações são disponibilizadas em *dashboard* onde é possível acompanhar diariamente a dispersão da ferrugem-asiática por estado, filtrar por municípios de ocorrência, ocorrências por datas e presença de esporos. Dessa forma, a utilização dos dados disponibilizados auxilia no controle da doença.

Considerando o impacto econômico ambiental de doenças em culturas de grande expressividade para o país e o potencial de gestão dessas endemias por um sistema de inteligência artificial, o presente trabalho visualizou a possibilidade de integração dessas duas problemáticas, neste sentido, a literatura apresentou um conjunto de pesquisas que contribuíram explicitamente para o desenvolvimento do presente trabalho. Megeto (2012), em seu trabalho executou e avaliou um processo de mineração de dados utilizando modelos computacionais, em algoritmos de árvore de decisão, para classificação de ocorrência de ferrugem asiática da soja, com base em variáveis climáticas de temperatura e precipitação. Ainda no mesmo trabalho, este autor demonstrou regras que permitiram identificar os principais fatores, dentre aqueles selecionados, que influenciam a ocorrência da ferrugem. O modelo preditivo concebido pelo autor apresentou taxa de acerto de 79,52% para predizer condições de ocorrência ou não ocorrência da doença.

O trabalho apresentado por Thamada (2016), para a cultura de café, considerando a ferrugem do cafeeiro, desenvolveu conjuntos estatísticos também gerados por mineração de dados e utilizando diversas técnicas de inteligência artificial para a análise de variáveis do ambiente agrícola, com objetivo de predizer aumentos na taxa de progresso da ferrugem do cafeeiro no período crítico da doença, também considerou dados climáticos e avaliações mensais da incidência da doença nas lavouras. Apresentou uma representação gráfica dessas regras, a fim de obter um modelo com maior equilíbrio, considerando a planta com alta e

baixa carga pendente de frutos – os conjuntos desenvolvidos mostraram acurácia acima de 70,00% para alta carga de doença (78,00% – limiar de 5 pontos percentuais (p.p). / 73,33% – 10 p.p.) e baixa carga (72,64% – 5 p.p.).

Pelo que foi apresentado anteriormente, as técnicas de aprendizado de máquina, alinhadas aos princípios da agricultura de precisão, podem prover soluções para os produtores agropecuários abrangendo o gerenciamento de riscos à produção, por meio de alertas climáticos e de predisposição à ocorrência de doenças e pragas nas plantações, entretanto, diversas são as possibilidades de combinação de algoritmos visando melhor acurácia no processo de predição. Considerando este um problema de pesquisa e, dada a importância da doença ferrugem asiática para a cultura da soja, bem como a disponibilidade de um sistema de informações público para esta doença, o presente trabalho tem por objetivo apresentar o desenvolvimento de técnicas de AM, utilizando, para tanto, a metodologia proposta por Megeto (2012), e a comparação de dois algoritmos – regressão logística e árvore de decisão, a partir de técnicas de agrupamentos de dados (clusterização) para predições de ocorrência da doença de ferrugem asiática na soja.

A metodologia proposta por Megeto (2012) utilizou dados abertos de estações meteorológicas na região dos municípios estudados, o que pode tornar produtos informacionais de agricultura de precisão mais baratos e acessíveis. A presente pesquisa adapta e simplifica a metodologia de Megeto, e utiliza ferramentas computacionais mais usuais ao ambiente de tecnologia da informação e inteligência artificial.

Pelo que foi apresentado, é possível perceber que o campo de estudo em agricultura de precisão é relativamente recente e os resultados decorrentes de pesquisas permitem um conjunto de benefícios, públicos e privados, sendo, portanto, fértil à investigação científica.

Importante salientar que ainda existe grande ocorrência da ferrugem asiática (*Phakopsora pachyrhizi*) na cultura de soja, uma vez que a variabilidade climática inter regional é elevada no tempo e no espaço, dificultando a generalização de prospecção de condições de ocorrência que podem vir a favorecer a aplicação racional do defensivo para controle da ferrugem na cultura da soja.

Diante desse contexto, este trabalho levanta a seguinte questão: as técnicas de aprendizado de máquina, aplicadas à interpretação de características climáticas que influenciam na ocorrência da doença, podem auxiliar na predição das condições de ocorrência da ferrugem asiática com acurácia suficiente à racionalização do uso de

agrotóxicos?

1.1 Objetivo Geral

Analisar a possibilidade do uso de aprendizado de máquina na previsão da ocorrência de ferrugem asiática na soja por meio de variáveis ambientais

1.1.1 Objetivos Específicos

- Aplicar diferentes ajustes de modelos preditivos para classificar a ocorrência da ferrugem da soja utilizando covariáveis ambientais;
- Comparar a acurácia dos modelos aplicados na predição de ocorrência da ferrugem asiática na soja a partir de covariáveis ambientais.

2. JUSTIFICATIVA

A 4ª Revolução Industrial levou ao campo Tecnologias da Informação, através do uso de sensores e dispositivos conectados, permitindo monitorar em tempo real, e de forma integrada, toda a cadeia produtiva rural, gerando uma gama imensurável de dados para análises. May (2010) classifica as tecnologias ambientais como um conjunto de conhecimentos, técnicas, métodos, processos, experiências e equipamentos que permitem utilizar os recursos naturais de forma sustentável e que possibilitam a disposição adequada dos rejeitos. Apesar de abrangente, podemos incluir soluções de aprendizado de máquina, como o modelo preditivo a ser analisado na presente pesquisa, dentro desta definição, uma vez que o mesmo permite compreender o cenário agroclimático que propicia menor incidência de fitopatologia na soja em determinada localidade, minimizando a aplicação de agrotóxicos e os efeitos danosos de seus rejeitos.

Porto e Soares (2012) destacam que os atuais processos de produção agrícola estão sendo criticados pelas ameaças à sustentabilidade ambiental e de saúde sendo urgente a demanda por alternativas sustentáveis. Diante desse cenário, o uso das novas tecnologias, como a inteligência artificial, podem transformar essa realidade, otimizando a cadeia

produtiva do agronegócio de maneira sustentável, ampliando a preservação ambiental, uma vez que o aumento da produtividade e a redução de custos pode gerar um efeito poupa terra.

As novas tecnologias aplicadas ao meio rural, possibilitam maior eficiência e competitividade, transformando o campo com precisão nos processos produtivos através da conectividade, tratamento, manipulação e processamento de dados que podem gerar informação para estratégias e ações.

Atualmente, a aplicação de fungicida para a ferrugem asiática é preventiva. Esse estudo permitirá a aplicação de fungicida baseado na possibilidade de ocorrência da ferrugem asiática em função das covariáveis ambientais utilizadas.

Portanto, o estudo justifica-se pela possibilidade de seu resultado proporcionar maior sustentabilidade à produção agropecuária, dado a possibilidade de redução do uso de fungicidas e pela melhoria na definição de estratégias para controle da doença na cultura da soja, mitigando os impactos ambientais. Ainda, a comunidade científica pode se beneficiar pelo incremento de trabalhos nesta temática, uma vez que valida os trabalhos já existentes e se diferencia dos demais pela proposta de comparação de desempenho entre algoritmos na ocorrência de ferrugem asiática.

3. REFERENCIAL TEÓRICO

3.1. Inteligência Artificial

A inteligência artificial ou inteligência computacional, é uma ciência recente, surge após a 2° guerra mundial, tendo seu nome evidenciado em 1956 (RIBEIRO, 2010). Segundo Ribeiro (2010, p.8), “a inteligência artificial é uma ciência multidisciplinar que busca desenvolver e aplicar técnicas computacionais que simulem o comportamento humano em atividades específicas”. Este campo de estudo busca encontrar técnicas e metodologias de desenvolvimento de software que aproximem a lógica humana e de máquina.

De acordo com Russel e Norvig (2013), os primeiros estudos sobre inteligência artificial datam de 1940 quando, durante a 2° Guerra Mundial, instigados pela inteligência britânica sobre a necessidade do desenvolvimento de métodos tecnológicos voltados principalmente a quebra de códigos criptografados e de cálculos para projetos de arma

nucleares, um grupo de matemáticos, liderados por Alan Turing, desenvolveram as primeiras pesquisas para criação de máquinas com vistas a realizar processamentos matemáticos de maneira mais rápida que a humana.

Outro importante estudo para o avanço na área da inteligência artificial data de 1943, com a publicação do trabalho de McCulloch e Pitts, em seu trabalho propuseram um modelo de neurônio artificial que se caracteriza por dois estados: ligado e desligado. O estado de ligado representava que o neurônio artificial estava recebendo um estímulo. A dupla de autores, com esse procedimento, conseguia adaptar qualquer problema de lógica computacional à rede de neurônios (RUSSELL; NORVIG, 2013).

Azevedo (2005) explica que existem três principais linhas de pesquisa na área da inteligência artificial: a linha conexionista, simbólica e a evolutiva. A linha conexionista propõe a modelagem da inteligência humana por meio de simulações dos neurônios e de suas interligações. Como exemplo é possível citar os estudos envolvendo redes neurais, tendo como marco histórico o desenvolvimento do Perceptron, no trabalho desenvolvido por Minsky e Papert em 1969, que posteriormente levou ao desenvolvimento das redes neurais artificiais (RUSSELL; NORVIG, 2013).

A linha simbólica utiliza formalismo do tipo lógico para simular o comportamento inteligente expresso através de linguagem. Nesta linha de pesquisa, buscava-se utilizar as máquinas para solucionar problemas de maneira a imitar o comportamento humano de divisão em sub-problemas para a resolução em pequenas partes que completariam o problema original de forma mais rápida.

A terceira linha de pesquisa da IA, a evolutiva, também chamada de computação evolutiva, se baseia na observação de mecanismos evolutivos encontrados na natureza, tais como a auto-organização e comportamento adaptativo. Os algoritmos genéticos e de autômatos celulares se destacam nessa linha de pesquisa, fomentando pesquisas na área de bioinformática. (RUSSELL; NORVIG, 2013).

3.2. Aprendizado de Máquina

Os avanços na área da inteligência artificial dividiram-se em subcampos, sendo um desses o Machine Learning, ou aprendizado de máquina (AM), que pode ser definido como: Ferramentas computacionais capazes de adquirir conhecimento de forma automática, que

toma decisões de experiências de problemas anteriores aprendendo novas regras ou inferências (MICHALSKI; CARBONELL; MITCHELL, 2013).

Trata-se do processo de examinar e experimentar, por meio de algoritmos que aprendem de maneira automática, a partir da experiência de um determinado fenômeno ou padrões de uma base de dados, possibilitando a definição de estratégias mais eficazes, por meio da previsão do comportamento dos resultados. Existem dois tipos de aprendizado de máquina: o aprendizado *não-supervisionado* e o *supervisionado.*

Aprendizado não-supervisionado, segundo Duda (2002), é o tipo de aprendizado em que o analista fornece apenas dados de entrada sem suas saídas correspondentes, desta forma o algoritmo fica a cargo de encontrar padrões nos dados. As técnicas de aprendizado não-supervisionadas são - agrupamento, associação e sumarização. Dentre estas, destaca-se a clusterização ou agrupamento, sendo considerada a mais importante deste tipo de aprendizado.

Ainda segundo Duda (2001), métodos de clusterização são definidos por algoritmos que determinam a divisão de grupos (clusters) distintos. A clusterização se fundamenta na ideia de distância ou similaridade dos dados para criar os grupos pertinentes, elementos que compõem um mesmo cluster devem apresentar uma alta similaridade entre si, uma vez que seguem um padrão similar de comportamento. Os algoritmos de clusterização podem ser usados como etapa de pré-processamento para outros algoritmos.

No aprendizado supervisionado, de acordo com Vieira (2016), os algoritmos utilizam para "treinamento" um conjunto prévio de dados coletados do mundo real. Esse conjunto de dados tem uma parte “x” que contém as características de interesse do problema, e outra parte, denominado de “y” que contém o valor das características de x ou a classe dos exemplos do mundo real. Assim, a ideia é que o sistema "aprenda" o mapeamento entre x e y a partir dos dados de treinamento, para que posteriormente seja capaz de "prever" o valor de y para um novo x, isto é, o valor de uma função, ou a classe à qual um novo exemplo pertence.

Ainda, Vieira (2016), explica que, se o domínio de Y for um conjunto ordenado de valores (p. ex., o conjunto dos números reais), o problema é dito de regressão. Porém, se o domínio de Y for um conjunto finito e não ordenado de valores, o problema é dito de classificação. Exemplo: estimar o preço de um carro mediante as características de potência do motor e tipo de material, seria um problema de regressão. Enquanto, determinar se uma

pessoa tem câncer benigno ou maligno, sabendo-se o tamanho, formato do tumor e idade do paciente, teríamos um problema de classificação.

Existem diversos métodos e técnicas para aprendizado de máquina supervisionado e não supervisionado, sendo que, neste trabalho foram apresentadas e utilizadas, árvores de decisão e regressão logística, dentre os algoritmos de aprendizado supervisionado, e dentre os algoritmos de aprendizado de máquina não supervisionado, optamos pelo uso da clusterização.

Segundo Garcia (2000), as Árvores de Decisão consistem de: (nós), que representam os atributos, e de (ramos), provenientes desses nós e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nós folhas (folha da árvore), que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

A partir disso, este algoritmo de classificação, através dessas variáveis e regras, consegue aprender padrões dentro de uma base de dados, sendo capaz de predizer, quando a pessoa irá jogar ou não jogar futebol, considerando as variáveis climáticas. Contudo, quanto mais variáveis e regras forem adicionadas ao algoritmo, mais robusto ele será. E a depender de quão bem seus dados, regras e modelo forem estruturados, maiores são as possibilidades de assertividade.

A regressão logística consoante com Hosmer & Lemeshow (2000), se caracteriza por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes, qualitativas ou métricas. O objetivo é refletir uma escolha entre duas alternativas, do tipo “sim ou não”, “isto ou aquilo”, “existe ou não existe”, respostas estas marcadas pelo seu caráter de alta objetividade. Os modelos mais simples dentro dessa classe são aqueles em que a escolha é uma ou outra entre (apenas) duas alternativas disponíveis e mutuamente exclusivas (OLIVEIRA, 1997).

De acordo com Gujarati (2000), nesses modelos, a variável dicotômica é uma variável *dummy*, assume os valores 0 ou 1, verifica a probabilidade de ocorrência do evento de interesse, ou seja, procura explicar determinados fenômenos quando a variável alvo é binária, por exemplo - “ocorrência” ou “não ocorrência” da ferrugem asiática.

Define-se P como a probabilidade da ocorrência da ferrugem asiática. Essa probabilidade pode ser determinada por vários fatores (variáveis explicativas), representadas

por X, voltadas sobre a variável dependente Y (ocorrência ou não ocorrência). Haja vista que a probabilidade é uma variável que se encontra no intervalo 0 e 1, a relação entre a probabilidade e determinado atributo pode ser representada. A que, na análise em questão, representa: 0 – Não ocorrência de ferrugem asiática, e 1 – ocorrência de ferrugem asiática.

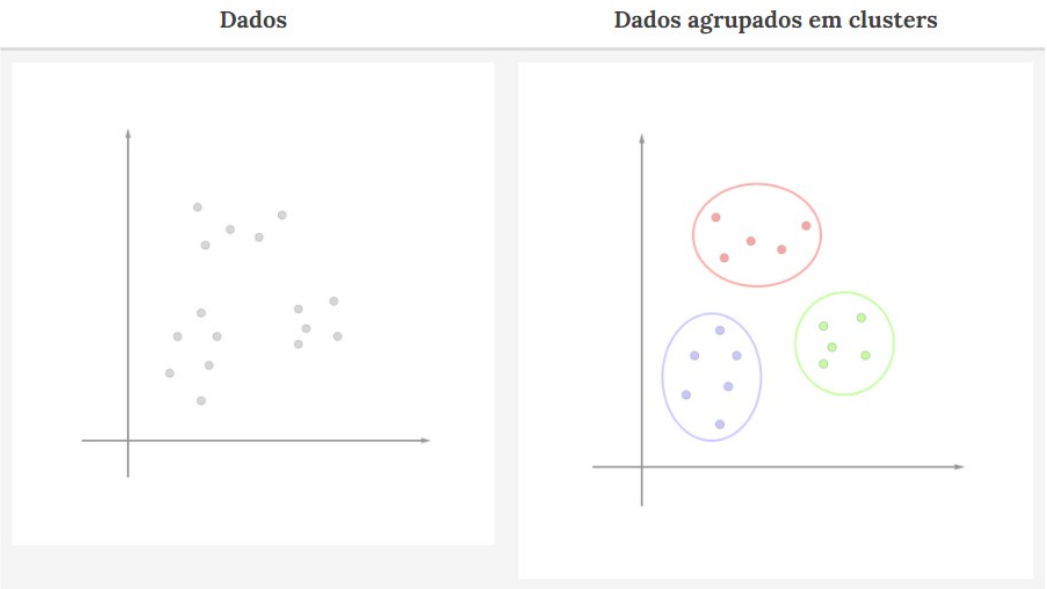
De maneira geral, a regressão logística atribui à variável resposta, um intervalo compreendido entre 0 e 1, que pode ser interpretado como a probabilidade de ocorrência de certo evento, propõe que baseado na regra de decisão, resultados superiores a 0,5 são assumidos como possível sucesso, e inferiores a esse valor, como possível fracasso.

Likas, Vlassis e Verbeek (2003) definem a clusterização como um algoritmo que dado um conjunto de dados, objetiva em uma quantidade X de subdivisões do conjunto de dados (clusters) formado a partir de uma amostra de centro (centróide) e demais dados cuja a distância euclidiana do ponto central estejam adjacentes. Fontana e Naldi (2009) descrevem o processo de clusterização, de maneira geral, o algoritmo K-means em quatro passos:

1. Atribuem-se valores iniciais para os centróides seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo;
2. Atribui-se cada objeto ao grupo cujo centróide possui maior similaridade com o objeto;
3. Recalcula-se o valor do centróide de cada grupo, como sendo a média dos objetos atuais do grupo;
4. Repete-se os passos 2 e 3 até que os grupos se estabilizarem;

Podemos observar o trabalho do algoritmo de clusterização (figura 1) em um determinado conjunto de dados antes e depois do agrupamento. O algoritmo organiza os dados em grupos de maior semelhança, o que torna a técnica adequada para problemas que envolvem dados não estruturados ou na preparação dos dados para a descoberta de padrões desconhecidos em dados estruturados. (LIKAS; VLASSIS; VERBEEK, 2003)

Figura 1. Conjunto de dados antes e depois da clusterização.



Fonte: www.lamfo-unb.github.io

3.3. Mineração de dados (Data Mining)

Segundo Figueira (1998) a tecnologia impulsiona a produção e acúmulo de dados, que são armazenados em bases de dados para uso futuro, a fim de identificar padrões e tendências. O volume massivo de dados produzidos é conhecido como big data, onde sistemas informatizados e sensores, quase que onipresentes, estão conectados gerando dados que são armazenados para garantir que possam ser recuperados futuramente por profissionais especializados, com objetivo de garantir a melhor decisão a ser tomada.

A mineração de dados, ou data mining, é o processo utilizado para extrair informações relevantes, muitas vezes ocultas em bases de dados, capaz de gerar conhecimento e contribuir para a solução de problemas nas organizações. Segundo Cabena et al. (1998) o processo de mineração de dados pode ser definido como um conjunto de técnicas interdisciplinares que combina áreas da Ciência da Computação, Inteligência Artificial e Estatística, sendo projetado para explorar massivas quantidades de dados à

procura de padrões que tenham valor para as organizações.

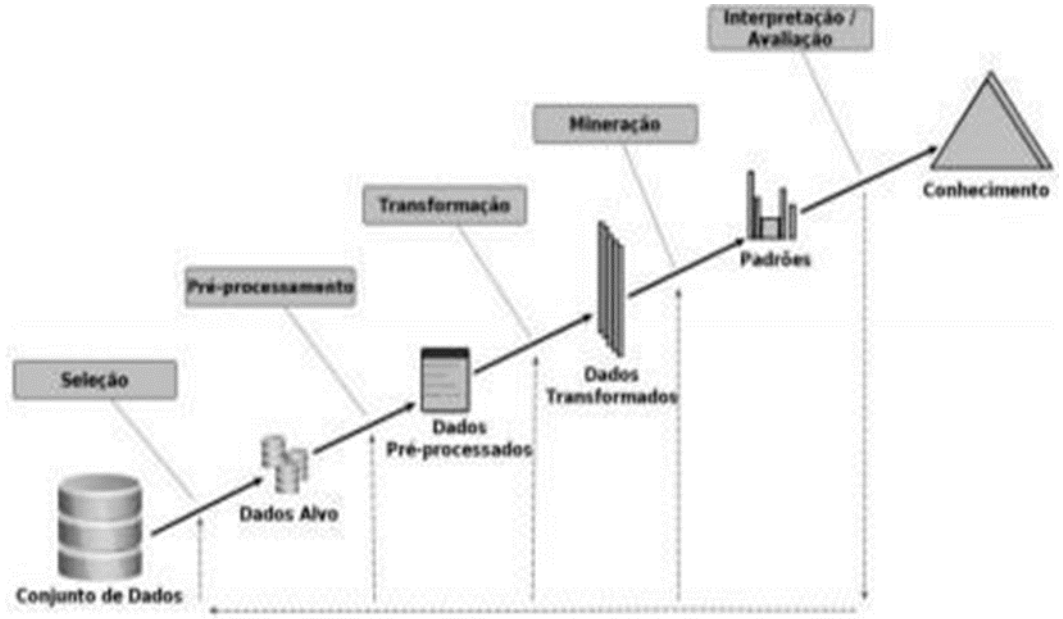
O processo de mineração de dados pode ser dividido em três etapas principais, pré-processamento – etapa responsável pela extração, integração, transformação, limpeza e seleção de dados (sendo nesta fase a maior concentração de esforço dos profissionais para garantir bons resultados); extração de padrões – etapa em que o analista escolhe, configura e executa um ou mais algoritmos (originários do *aprendizado de máquina*); pós processamento

– consiste na adequação do algoritmo, com o objetivo de avaliar e validar se o resultado obtido é significante, ou se novo padrão e/ou tendência foram identificados. De acordo com Amaral (2016) a mineração de dados, processo conhecido como KDD – Knowledge Discovery Database (Descoberta de Conhecimento em Banco de Dados) – (figura 2), pode ser dividida em cinco fases:

- Entendimento do Negócio – consiste em compreender as características do negócio em que a mineração será aplicada;
- Pré-processamento – avaliar os dados em termos de estrutura, relacionamentos, qualidade, quantidade e acesso;
- Transformação – envolve tarefas como padronização de dados, dados organizados, limpos, selecionados para utilização de algoritmo de aprendizado de máquina;
- Mineração de dados – consiste na construção do modelo através do aprendizado de máquina, o modelo será utilizado para classificar dados ainda não conhecidos por ele;
- Interpretação e Avaliação – o modelo é avaliado por seu desempenho, posteriormente ao processo de mineração e é implantado.

A mineração de dados é comumente considerada como sinônimo de KDD, sendo utilizada para explorar grandes quantidades de dados, seja relacionado a mercado, negócios e pesquisas acadêmicas com o objetivo de encontrar/descobrir padrões consistentes que podem revelar informações estratégicas que agregam valor ao processo decisório de uma organização, como por exemplo, a elaboração de modelos preditivos para gerar alertas para o agronegócio referente a doenças, pragas, secas e geadas.

Figura 2. Processo KDD.



(Adaptado de Fayyad et al. (1996))

3.4. Ferrugem Asiática (*Phakopsora pachyrhizi*)

A ferrugem asiática, causada pelo fungo *Phakopsora pachyrhizi* Sydow e Sydow, foi descrita pela primeira vez no Japão em 1902 (Hennings, 1903). No ano de 1914 já havia se disseminado para diversos países do sudeste da Ásia. Nas Américas, essa doença foi relatada pela primeira vez no ano de 1976 em Porto Rico (Vakili e Bromfield, 1976), sendo reportada no Brasil a partir de 2002 (COSTAMILAN, ET AL. 2002; REIS ET AL., 2002). Atualmente a ferrugem está presente em todos os países que cultivam soja.

Os primeiros sintomas da ferrugem da soja causada por Phakopsora pachyrhizi se caracterizam por pequenas manchas com coloração que pode variar do cinza esverdeado ao marrom-avermelhado (Hartman et al., 1999). Esses sintomas podem aparecer em qualquer estádio de desenvolvimento da planta e em diferentes tecidos da planta, como nas folhas e hastes (Almeida et al., 2005).

Consoante com Reis et al (2006), no campo, estas manchas inicialmente aparecem nas folhas mais baixas do dossel durante ou após a floração, embora as plântulas podem ser infectadas sob certas circunstâncias. As primeiras lesões aparecem na base do folíolo perto do pecíolo e das nervuras da folha e permanecem pequenas (2-5 mm de diâmetro). Com o

progresso da doença as lesões aumentam de tamanho (Reis et al., 2006a; Sinclair e Backman, 1993).

Segundo Almeida et al (2005), as urédias ou pústulas encontradas nas lesões, são os corpos de frutificação dos fungos causadores das ferrugens, elas produzem uma grande quantidade de uredósporos, que são os esporos responsáveis pela disseminação da doença no campo (figura 3). As urédias maduras podem ser vistas a olho nu, especialmente durante a esporulação. Mesmo uma pequena lesão pode possuir uma grande quantidade de urédias e consequentemente, de esporos infectivos, o que proporciona a disseminação da doença entre as plantas (Hartwig, 1986; Kochman, 1977; Bonde et al., 2006; Miles et al., 2006).

Uma vez que os esporos viáveis se depositam na superfície de um hospedeiro suscetível, a infecção e o desenvolvimento da epidemia são dependentes das condições ambientais. Para Nicolini et al. (2010), geralmente, as condições favoráveis para o sucesso da infecção são folhas úmidas e temperaturas entre 8°C e 28°C, sendo que os limites ótimos de 16°C a 28°C.

Figura 3. Folhas de soja infectada com o fungo Phakopsora pachyrhizi, causador da ferrugem asiática da soja, ilustrando as lesões resultantes da infecção.

Fonte: www.agencia.cnptia.embrapa.br

Segundo Reis et al. (2006), após a infecção, lesões e pústulas com uredósporos podem aparecer em 7 ou 8 dias, quando então o próximo ciclo de infecção pode ter início,

ou seja, sob condições favoráveis as epidemias de ferrugem da soja podem progredir rapidamente de baixos níveis de detecção para desfolha da planta dentro de um curto espaço de tempo. A idade das plantas também pode afetar as epidemias de ferrugem da soja, sendo que geralmente, as lesões de ferrugem não são encontradas antes do florescimento da soja, a não ser que se tenham altos níveis de inóculo no início da estação.

Ainda consoante Reis et al. (2006), com o progresso da doença, as áreas infectadas começam a amarelar e, eventualmente, as plantas começam a perder as folhas. Apesar dos primeiros sintomas aparecerem nas partes mais baixas do dossel, a ferrugem avança rapidamente para as folhas mais altas até que todas as folhas apresentem os sintomas da doença. A perda desses tecidos foliares resulta na redução de produtividade devido à diminuição do tamanho e número de grãos. Danos na produtividade de 30 a 80% já foram relatados, porém o volume dos danos depende do estádio no qual a doença iniciou e da velocidade de progressão.

A ferrugem asiática da soja tem sido uma ameaça fitossanitária permanente, com agravamento nas últimas safras, decorrente do fenômeno da resistência do fungo e, em consequência da diminuição de eficiência de fungicidas utilizados desde o surgimento da doença. O uso de fungicidas consiste em umas das principais formas de preservar a eficiência de controle da ferrugem asiática. Recentemente na de safra 2016/17, os fungicidas inibidores apresentaram redução de eficiência, por esse motivo, práticas de manejo integrado, que incluem o uso de fungicidas com outros modos de ação, são atualmente recomendadas. O manejo da ferrugem asiática da soja deve ser adotado de forma integrada, baseado em três principais medidas de controle, apresentadas abaixo (Godoy e Meyer, 2017):

1. Controle químico através do uso de fungicidas foliares;
2. Adoção de cultivares resistentes e semeadura de cultivares precoces no início da época de recomendada para cada região;
3. Medidas legislativas instituindo o vazio sanitário e a calendarização de cultivo da soja.

Rupe et al. (2008) mencionam o uso de parcelas sentinelas como forma de previsão e avaliação da gravidade com que a ferrugem asiática pode infectar a lavoura. Cultivares de ciclo mais curto semeados por volta de 03 semanas antes do plantio comercial alcançarão o estágio de floração antes da cultura comercial, permitindo os produtores aplicarem o tratamento de fungicida.

Outro método mencionado por Rupe et al. (2008) é a captura de esporos, seja utilizando placas de vidro para capturar os esporos presentes no vento, ou com coleta e filtragem da água da chuva. Em ambas as técnicas, é necessário o apoio da avaliação dos esporos capturados em laboratório com o uso de microscópios.

4. METODOLOGIA

A presente pesquisa se caracteriza pela natureza aplicada, pois objetiva gerar conhecimento para aplicação prática dirigida a um problema específico. O modelo resultante do trabalho, assim como os procedimentos do tratamento de dados, poderão ser utilizados em outras pesquisas que utilizam variáveis climáticas como elementos de categorização de doenças. Quanto a forma de abordagem, caracteriza-se como exploratória e quantitativa, uma vez que procura obter maior familiaridade com o problema (GIL 2002), relacionado aos valores das características climáticas que levam a ocorrência da ferrugem asiática da soja.

Relativo aos objetivos e procedimentos técnicos do trabalho trata-se de uma pesquisa experimental, ao executar experimentos (testes) com variáveis climáticas que influenciam a ocorrência do fenômeno (GIL 2002). A construção do modelo preditivo da ferrugem asiática da soja foi definido a partir da comparação dos resultados obtidos entre os dois algoritmos de aprendizado de máquina, árvore de decisão (A) e o algoritmo de regressão logística (B).

Este estudo adaptou a metodologia proposta por Megeto (2012) para coleta de dados e definição da ocorrência da ferrugem asiática e também a metodologia KDD (seção 3.4) para avaliação do desempenho dos algoritmos. O tratamento dos dados e os modelos foram elaborados na linguagem Python versão 3.9 (Van Rossum and Drake, 2009), desenvolvido sob uma licença de código aberto e livre para ser utilizável e distribuível, amplamente utilizado por comunidade de desenvolvedores e acadêmica, por sua flexibilidade.

Para a construção dos modelos foram utilizadas três bases de dados, sendo os atributos climáticos provenientes dos portais Agritempo (Sistema de Monitoramento Agrometeorológico) e INMET (Instituto Nacional de Meteorologia), e os dados de registros de ocorrência de ferrugem asiática da soja, obtidos no site Consórcio Anti-Ferrugem (CAF), todos os portais com dados abertos, disponíveis para download de arquivos no formato ".csv" (*comma separated values* – valores separados por vírgulas).

O Agritempo consiste em um sistema web que disponibiliza, gratuitamente na internet, informações e dados agrometeorológicos com cobertura para todo o território brasileiro, é alimentado por dados provenientes de mais de 1.400 estações meteorológicas de superfície, o sistema possui uma série de funcionalidades como o download de séries históricas por cidade e suas estações meteorológicas, sendo os dados de temperaturas mínima, média, máxima e precipitação oriundos do Agritempo, e os dados de radiação solar

e umidade do INMET (Agritempo, 2021).

O INMET é um Instituto do Ministério da Agricultura, Pecuária e Abastecimento, e tem por objetivo fornecer informações meteorológicas à sociedade brasileira que possam contribuir no processo de tomada de decisão e para o desenvolvimento sustentável do País. O INMET disponibiliza em nível nacional, a previsão do tempo, avisos e boletins meteorológicos especiais, o Instituto conta com um sistema de coleta e distribuição de dados meteorológicos de temperatura, umidade relativa do ar, direção e velocidade do vento, pressão atmosférica, precipitação, entre outras variáveis, possui estações de sondagem de ar superior (radiossonda); estações meteorológicas de superfície sendo a maior rede de estações automáticas da América do Sul (INMET, 2021).

O Consórcio Antiferrugem foi criado em 2004, como uma estratégia para divulgar informações sobre a ferrugem asiática da soja, possui cerca de 60 pesquisadores de instituições públicas e privadas, com o intuito de monitorar o problema e gerar informação atualizada sobre a doença, conta com aproximadamente 100 laboratórios cadastrados e disponibiliza registros de ocorrência de ferrugem asiática nos municípios do Brasil (Embrapa, 2021).

Fundamentadas as variáveis climáticas conforme revisão sobre a temática Ferrugem Asiática, e Megeto (2012), foram selecionadas as **seguintes variáveis:**

- **Temperatura**: grandeza que caracteriza o calor, (agitação molecular) de um corpo ou sistema.
- **Umidade relativa do ar**: representa a quantidade de água presente no ar em relação à capacidade de saturação do vapor na atmosfera.
- **Precipitação Pluviométrica**: Processo pelo qual a água condensada na atmosfera atinge gravitacionalmente a superfície terrestre, gerado por meio do aumento da pressão do vapor de água na atmosfera ou pelo resfriamento do ar úmido.
- **Radiação Solar global**: energia emitida pelo sol e transmitida sob a forma de ondas eletromagnéticas. Cerca de metade dessa energia é emitida na forma de luz visível, e a outra metade em ondas infravermelho e ultravioleta

4.1. Área de estudo

Conforme a definição das variáveis, foi realizado o download da série histórica climática de 2006 à 2017, referente aos municípios de Cristalina (GO), Jataí (GO) e Rio Verde (GO), que segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), estão entre os dez municípios com maior valor de produção agrícola do Brasil. O período de 2006 a 2017 foi escolhido devido a disponibilidade de dados climáticos e ocorrências de ferrugem asiática confirmadas. Os registros de ocorrência da ferrugem asiática referente ao mesmo período, foram retirados do site Consórcio Antiferrugem (CONSÓRCIO ANTIFERRUGEM, 2021), que disponibiliza informações como: data da ocorrência, safra, município, momento da detecção, mês de plantio e quinzena de plantio.

4.2. Procedimentos de Pesquisa

A pesquisa consiste em quatro etapas, realizadas da seguinte maneira:

- **Etapa I**: coleta e tratamento dos dados – cada registro coletado de ocorrência da doença é uma ocorrência verdadeira que possui confirmação por laboratório especializado, assim, atribui-se o atributo "Ocorrência – OC". Para cada ocorrência cria-se um registro de "Não Ocorrência – NOC" definido como o trigésimo dia anterior ao dia de ocorrência da doença.

Segundo Megeto (2012) a NÃO OCORRÊNCIA deve ser estabelecida com o objetivo de assegurar que exista distinção entre os fatores climáticos que possam favorecer a ocorrência, desta forma, "desenvolveu-se um procedimento para simular um evento de não ocorrência da doença. Para cada registro de ocorrência foi criado um evento de não ocorrência, com a premissa de que em um momento anterior, no caso 30 dias antes da ocorrência, a doença não havia sido detectada" (MEGETO, 2012, p.31).

A partir dos dias de ocorrência e não ocorrência foi gerado um único arquivo em formato ".csv" totalizando 393 linhas contendo registros de ocorrência e tendo nas colunas atributos climáticos referentes às ocorrências (tabela 1), constituindo assim uma tabela final com todos atributos relacionados à cada ocorrência.

O conjunto de atributos climáticos que constituíram o arquivo final em ".csv" precisaram de tratamento devido dados faltantes nas variáveis Radiação Solar, Umidade e Precipitação, inicialmente seria usada somente a base de dados do Agritempo, entretanto diante da grande quantidade de dados faltantes para as variáveis acima citadas, optou-se por

recorrer a base do INMET, que oferece dados mais consistentes por realizarem medidas a cada 3 horas, totalizando 8 medições diárias, porém nas cidades de Cristalina e Jataí, a série histórica continha somente registros até o ano de 2013, sendo necessário integrar as duas bases climáticas, e assim evitar maior tratamento de dados por parte dos pesquisadores, garantindo uma representação mais fidedigna da realidade referente aos dados climáticos.

Tabela 1. Atributos de cada ocorrência contida na tabela utilizada na pesquisa.

Atributos	Descrição	Exemplo
Data	dd/mm/aaaa	06/02/2008
Cidade	Nome das cidades por extenso	Rio Verde
Ocorrencia_Registrada	Categórico: Sim / Não	Sim
Ocorrencias_por_dia	Numérico: Quantidade de ocorrência no dia	36
Precipitacao(mm)	Numérico: Precipitação acumulada do dia, em milímetros (mm)	176,5
RadSol(MJ/m2)	Numérico: Radiação Solar acumulada no dia (MJ/m²)	4,31
Temp_Med	Numérico: Temperatura média do dia, em graus Celsius (ºC)	19,31
Temp_Max	Numérico: Temperatura máxima do dia, em graus Celsius (ºC)	25,38
Temp_Min	Numérico: Temperatura mínima do dia, em graus Celsius (ºC)	16,69
UmidRel(%)	Numérico: Média de Umidade Relativa do Ar (%)	63,52

Fonte: Elaborada pelos autores.

Em um segundo momento, após a integração das duas bases climáticas, os anos de 2013 e 2014 persistiram com dados faltantes nas variáveis Umidade e Radiação Solar, e para resolver como as lacunas de dados seriam preenchidas, elaboramos a estratégia de realizar a média dos 5 dias anteriores dos dados existentes e dessa forma usar essa média para preencher os dados inexistentes, alcançando um de nossos objetivos específicos referente ao tratamento da base de dados.

- **Etapa II**: implementação de algoritmos de aprendizado de máquina A e B – para a implementação dos algoritmos na linguagem python, foi utilizada a interface livre de código aberto *Jupyter Notebook.* Nesta pesquisa, conforme citado anteriormente, foram utilizadas as bibliotecas *python* própria para desenvolvimento em *machine learning*, *scikit-learn,* conforme apresentado por Pedrosa et al. (2011).

Esta biblioteca contém os algoritmos de árvore de decisão e regressão logística, avaliados nesta pesquisa, assim como funções de pré-processamento de dados que também foram utilizadas no desenvolvimento deste trabalho. Os parâmetros de cada algoritmo foram salvos, junto com o seu resultado, para posterior avaliação de desempenho e definição do modelo final de classificação de ferrugem asiática.

- **Etapa III**: avaliação de desempenho dos algoritmos utilizados – os testes foram realizados com a metodologia de validação *K-fold cross-validation* – validação cruzada de k pastas. Esta validação consiste em dividir aleatoriamente o conjunto de dados em k pastas, onde k-1 pastas serão utilizadas para treinamento do modelo, e a pasta restante é utilizada para testar o modelo.

Treinamentos e testes foram realizados até o ponto em que houve "k" testes distintos (REFAEILZADEH et al., 2008).

Para os algoritmos regressão logística e árvores de decisão existem métricas e critérios para que, subsequente a elaboração do modelo, o desempenho seja analisado (Kamber 2001). O principal critério de avaliação considerado nesta pesquisa foi o de acurácia (precisão), definido como a capacidade do modelo em determinar rótulos de classes corretos para instâncias novas ou desconhecidas, ou seja, a capacidade de predição correta do classificador.

Após cada conjunto de validações, os resultados foram avaliados e os parâmetros dos algoritmos (tabela 2.1 e 2.2) foram modificados e os resultados novamente avaliados (acurácia).

Tabela 2. Parâmetros alterados no modelo de regressão logística.

Regressão Logística	
Parâmetro	Definição
Random_State	Número utilizado pelo gerador de valores aleatórios para randomizar os dados de treinamento
Max_Iter	Número máximo de iterações de treinamento
Tol	Tolerância para o critério de parada do treinamento

Fonte: Elaborada pelos autores

Tabela 3. Parâmetros alterados no modelo de árvore de decisão.

Árvore de Decisão	
Parâmetro	Definição
Criterion	A função que mede a qualidade da divisão de nós. As funções disponíveis são: “Gini” referente a impureza Gini e “Entropy” para ganho de informação
Random_State	Número utilizado pelo gerador de valores aleatórios para randomizar os dados de treinamento
Min_samples_Leaf	Número mínimo de amostras que formarão uma folha nó
Min_samples_Split	Número mínimo de amostras necessárias para separar um nó
Presort	Pré organização dos dados para agilizar o processo de definição das melhores divisões dos nós

Fonte: Elaborada pelos autore

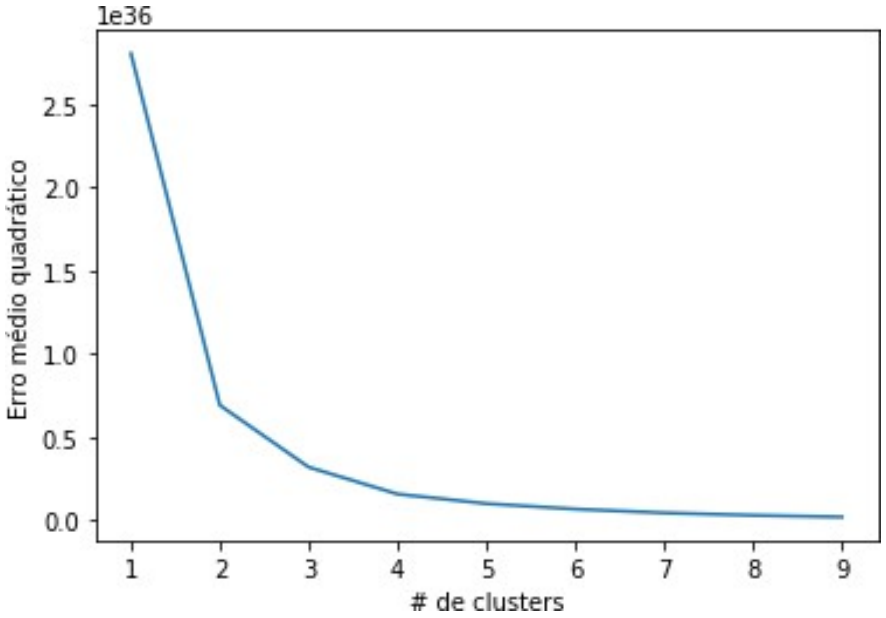
- **Etapa IV:** Clusterização dos dados - Após a avaliação dos resultados obtidos nos algoritmos, utilizando a acurácia como métrica de avaliação, os processos descritos nas etapas II e III foram retomados utilizando o algoritmo de clusterização k-means como forma de preparação de dados.

Os algoritmos de clusterização agrupam dados de determinada similaridade, entretanto, os dados desta pesquisa apresentaram elevada similaridade, em termos de amplitude de variação, nos atributos relacionados ao clima do Cerrado (precipitação, radiação solar, temperaturas e umidade relativa) durante todo o período de safra da soja, ou seja, nos períodos de ocorrência e não ocorrência.

A fim de definir a quantidade ótima de grupos (clusters), foi utilizado o teste do “cotovelo”, recomendado conforme as boas práticas do algoritmo de clusterização (Bholowalia e Kumar 2014). O teste constitui-se em uma análise gráfica que demonstra em qual quantidade de clusters reside uma queda brusca (cotovelo) do erro médio quadrático. Para tanto, foram considerados apenas os atributos climáticos (precipitação, radiação solar, temperaturas e umidade relativa) como variáveis independentes, os quais foram regredidos ao atributo de classificação, como variável dependente.

Conforme observado (figura 4) o teste indicou o número de clusters em 02 como a redução drástica do erro médio quadrático.

Figura 4. Redução do erro médio quadrático conforme o número de clusters - “Teste do cotovelo”.



Fonte: Resultados da pesquisa

O conjunto de dados, após a utilização do algoritmo de clusterização, passa a conter uma nova coluna, que representa a qual agrupamento de dados aquela amostra se refere. Estes dados, agora agrupados, foram utilizados como entrada para execução dos modelos que já haviam sido avaliados neste trabalho.

5. APRESENTAÇÃO E DISCUSSÃO RESULTADOS

5.1. Análise Exploratória

As ocorrências da ferrugem asiática foram caracterizadas conforme local de ocorrência (tabela 4). A partir das medidas estatistísticas da base de dados (tabela 5 e 6), verificamos que entre os dados de ocorrência e não ocorrência não há grande variação nos dados climáticos dos municípios estudados

Tabela 4. Total de ocorrências por município (2006/2017).

Cidade	Ocorrências
Rio Verde	334
Jataí	141
Cristalina	16
Total	**491**

Fonte: Resultados da pesquisa

Tabela 5. Medidas Estatísticas de Ocorrências (2006/2017).

Medidas Estatísticas de Ocorrências				
	Precipitação	Temperatura Média	Radiação Solar	Umidade Relativa
Média	35,50	24,49	11,88	76,19
Desvio padrão	62,15	1,72	8,75	8,16
Mínimo	0.0	14,00	0,10	53,75
25%	1,20	23,65	3,83	71,00
Mediana	9,25	24,73	11,42	75,50
75%	34,54	25,55	19,88	81,50
Máximo	286,25	28,50	30,13	98,75

Fonte: Resultados da pesquisa

Tabela 6. Medidas Estatísticas de Não Ocorrências (2006/2017).

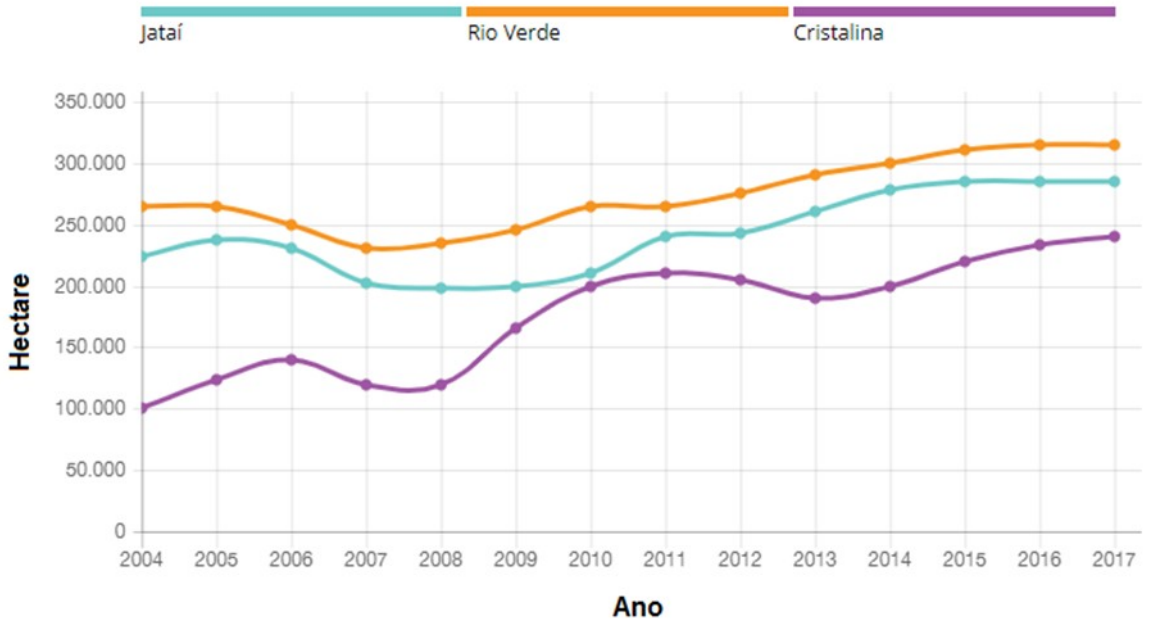
Medidas Estatísticas de Não Ocorrências				
	Precipitação	Temperatura Média	Radiação Solar	Umidade Relativa
Média	42,65	24,75	11,29	74,27
Desvio padrão	79,20	1,99	9,33	8,68
Mínimo	0,00	16,75	0,10	50,00
25%	0,20	23,70	3,10	68,50
Mediana	5,30	25,00	7,50	74,50
75%	64,40	26,00	20,36	78,75
Máximo	510,75	30,00	30,57	100,00

Fonte: Resultados da pesquisa

A ocorrência da ferrugem asiática na série histórica analisada (2006 - 2017) foi mais frequente na cidade de Rio Verde, seguida de Jataí, sendo consideravelmente menor em Cristalina.

Tal fato pode ser explicado pela quantidade de área plantada (ha) com soja ao longo do período analisado nos municípios em questão (figura 5), uma vez que esta quantidade é proporcional ao nível de ocorrências nos municípios.

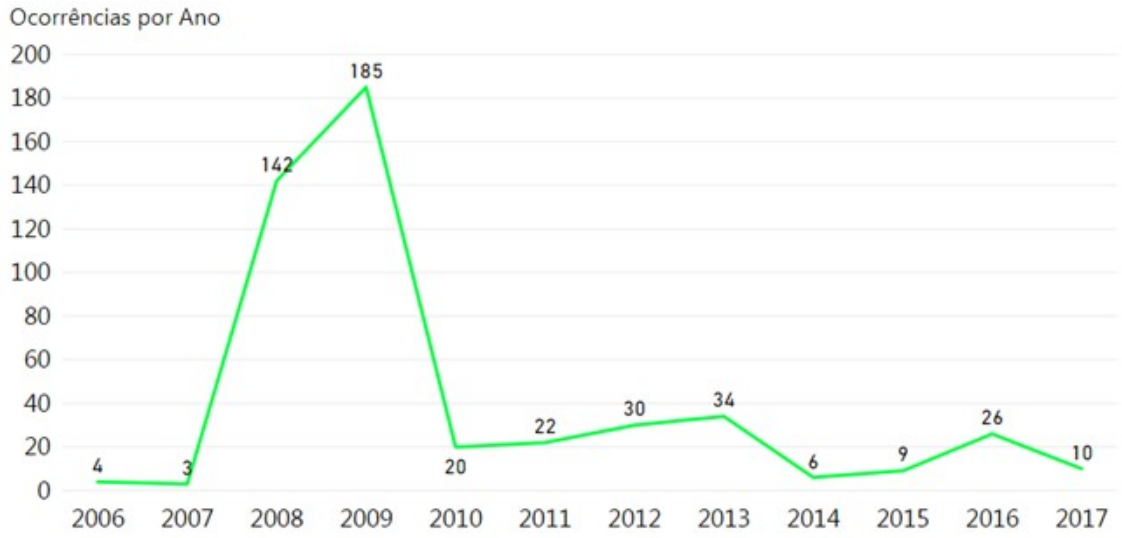
Figura 5. Área plantada com soja (2004 – 2017) nos municípios analisados pela pesquisa.



Fonte: IBGE/SIDRA (2020)

Entretanto, apesar de a área de soja plantada aumentar ao longo dos anos, o total geral de ocorrências registradas (figura 6) não reflete esse mesmo padrão, uma vez que, nos anos de 2008 e 2009, foram acentuadamente maiores os números de ocorrências registradas.

Figura 6. Total geral de ocorrências por ano.



Fonte: Resultados da pesquisa.

Segundo Marques (2011) este fato normalmente está associado a uma falha gerencial dos produtores, devido ao aumento dos custos de combate à doença e à redução da eficácia de alguns fungicidas. Tal constatação pode ser evidenciada quando comparada ao ano agrícola de 2014/2015, quando o combate à ferrugem ocorreu de forma mais intensa, segundo o portal Consórcio Antiferrugem (CAF, 2018).

A fim de observar a intensidade das correlações entre os atributos considerados no modelo (tabela 1) e as ocorrências registradas de ferrugem asiática na soja, foi gerada uma matriz de correlação (figura 7). As correlações existentes entre qualquer uma das variáveis (atributos) foram de fraca intensidade. Apesar da baixa correlação, esta não influencia no modelo de predição, uma vez que o mesmo leva em consideração a combinação desses atributos como informação relevante, além disso, a baixa correlação pode ser devido ao fato de que a amplitude de valores para os mesmos estar dentro da faixa de maior probabilidade de ocorrência da doença.

Figura 7. Matriz de correlação de ocorrência com variáveis climáticas - Captura de tela Jupyter Notebook.

	Data	Ocorrência Registrada	Ocorrencias por dia
Data	1.000000	-0.002261	-0.112940
Ocorrência Registrada	-0.002261	1.000000	0.457648
Ocorrencias por dia	-0.112940	0.457648	1.000000
Precipitacao(mm)	-0.182041	-0.050471	-0.045267
RadSol(MJ/m2)	0.070613	0.032635	0.044447
Temp_Med	0.122532	-0.070466	-0.094557
Temp_Max	0.010781	-0.075620	-0.056927
Temp_Min	0.202353	-0.023042	-0.101417
UmidRel(%)	0.125234	0.114059	0.053548

Fonte: Resultados da pesquisa

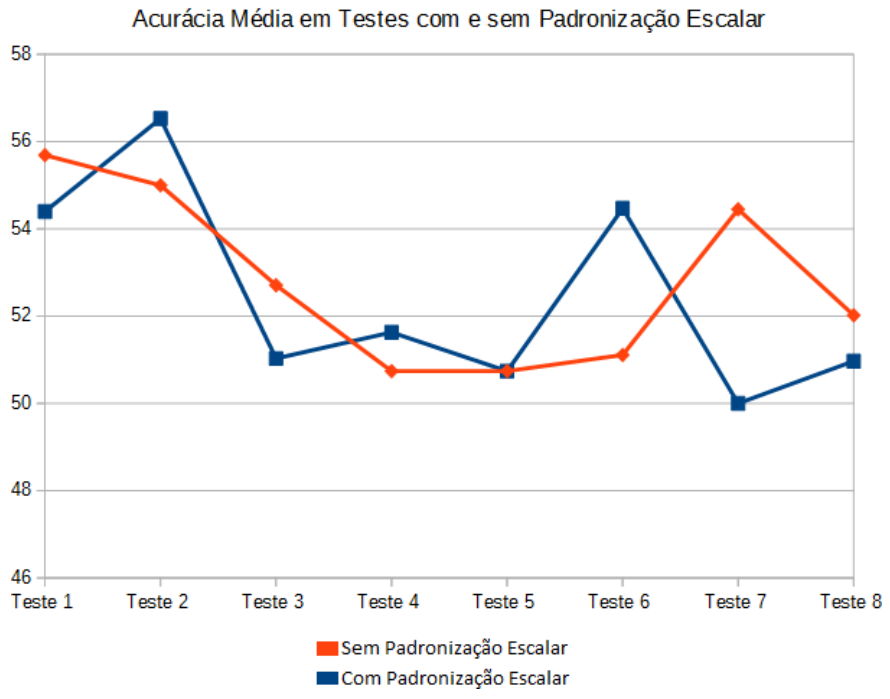
5.2. Regressão Logística

Foram realizadas 8 rodadas de avaliação e ajuste de parâmetros, entre estas, novos valores de parâmetros foram definidos para o modelo. Os testes foram divididos em 8 sem nenhuma padronização e 8 testes com padronização dos dados, realizada por meio da

função *StandardScaler*, (responsável por colocar os valores numéricos em escala equivalente), totalizando 16 testes para cada algoritmo. Para permitir uma avaliação comparativa, foram utilizadas as mesmas amostras em testes com e sem padronização escalar de dados.

No geral os modelos gerados com o algoritmo de regressão logística obtiveram apenas 56,53%, 54,47% e 54,45% em seus melhores resultados médios para cada valor de k, com e sem padronização escalar (figura 8) .

Figura 8. Acurácia média em testes de regressão logística com e sem padronização escalar.



Fonte: Resultados da pesquisa

Os testes dos modelos gerados a partir do algoritmo de regressão logística variaram em 3 valores de k para a validação cruzada: 10, 30 e 29 (tabelas 7, 8 e 9), dessa forma o modelo realizou k testes com 29, 9 e 10 amostras para validação respectivamente.

Com a finalidade de avaliar o significado das datas de ocorrência para a classificação da ferrugem asiática,o modelo de regressão logística testado inicialmente não

considerou este atributo nas classificações de ocorrências da doença, além disso, os resultados distantes da acurácia média não foram considerados válidos, pois além de assertivo, o modelo deve ser consistente para seu uso posterior.

Tabela 7. Resultados dos testes de regressão logística com k = 10.

Validações	Ajuste 1		Ajuste 2		Ajuste 3	
	Padronização Escalar		Padronização Escalar		Padronização Escalar	
	Não	Sim	Não	Sim	Não	Sim
Acurácia Média	**55,69**	**54,4**	**55**	**56,53**	**52,71**	**51,03**

Fonte: Resultados da pesquisa

O melhor resultado dos testes apresentados para 10 pastas de dados (tabela 7) foi de 56,53% de acurácia média em teste com padronização escalar. Os parâmetros para encontrar o melhor resultado destes testes foram: Random_State em 101 e Max_Iter em 150 iterações.

Os demais ajustes (tabelas 8 e 9) incluiram as datas da ocorrência ou não ocorrência da ferrugem asiática no modelo de predição.

Com o valor de k ajustado para 30 pastas, o modelo incluiu 284 amostras para treinamento e 9 para cada validação. O melhor resultado obtido nos testes com 30 pastas (tabela 8) alcançou uma acurácia média de 54,47% em teste sem padronização escalar, com parâmetros Max_Iter realizando até 1000 iterações por treinamento, e Tol com valor 1. resultado inferior aos testes com 10 divisões dos dados (tabela 8), além disso, a classe de não ocorrência era menos assertiva que a classe de ocorrência.

Tabela 8. Resultados dos testes de regressão logística com k = 30.

Validações	Ajuste 4		Ajuste 5		Ajuste 6	
	Padronização Escalar		Padronização Escalar		Padronização Escalar	
	Não	Sim	Sim	Não	Não	Sim
Acurácia Média	**51,63**	**50,74**	**50,74**	**50,74**	**54,47**	**51,11**

Fonte: Resultados do modelo

Finalmente foi conduzido um teste com 29 pastas, com cada rodada de 10 amostras

para validação, com a finalidade de comparar o significado da amplitude de variação da quantidade de amostras retidas para validação do modelo (tabela 9).

Tabela 9. Resultados dos testes de regressão logística com k = 29.

Validações	Ajuste 7		Ajuste 8	
	Padronização Escalar		Padronização Escalar	
	Não	Sim	Sim	Não
Acurácia Média	**50**	**54,45**	**52,02**	**50,97**

Fonte: Resultados da pesquisa.

Com 54,45% de precisão média em teste realizado com padronização escalar dos dados (tabela 9) foi possível concluir que, dentre os parâmetros testado, o resultado de acurácia para regressão logística foi mais sensível a quantidade de amostras retidas para validação que ao número de pastas de validação cruzada, além disso, com uma acurácia próxima de 50%, este modelo não seria adequado para previsão de doenças uma vez que não conseguiu compreender a diferença entre as combinações de variáveis climáticas entre dias de ocorrência e dias de não ocorrência, dado seu desempenho na classificação de cada classe.

A título de ilustração, foram gerados os valores das métricas de avalição para a sexta rodada de teste (ajuste 6) sem padronização dos dados, quando houve a melhor média de testes com k = 30 (tabela 8).

Os resultados (tabela 10) demonstram que os modelos gerados a partir do algoritmo de regressão logística não classificaram com assertividade a classe de não ocorrência, para esta classe o resultado do algoritmo alcançou recall (medida das amostras classificadas corretamente sobre o total de amostras da classe) e f1-score (relação entre precisão e recall) com 0,00% de precisão para a classe de não ocorrência. Uma das possibilidades deste fenômeno pode ser atribuída a pouca clareza na diferença dos dados climáticos de ambas as classes, para corrigir diferença, o algoritmo atribui ou “força” a classe positiva para as amostras.

Tabela 10. Métricas amostra k=30.

Classe	Precisão	Recall	F1-Score	Amostras
Não Ocorrência (0)	0,00%	0,00%	0,00%	34
Ocorrência (1)	54,00%	100,00%	70,00%	40
Média por classe	27,00%	50,00%	35,00%	74

Fonte: Resultados da pesquisa.

5.3. Árvore De Decisão

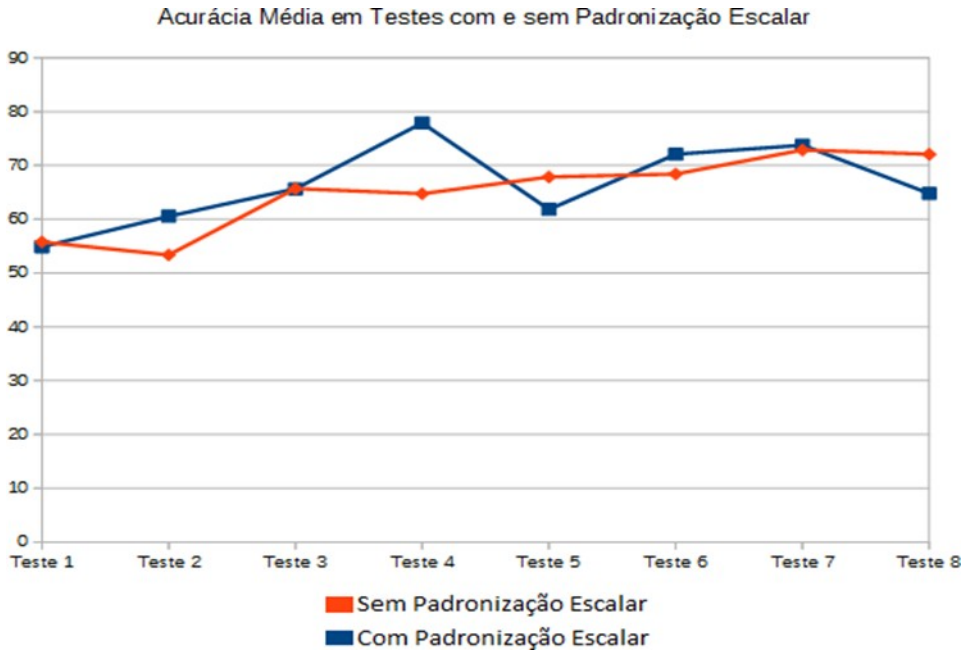
Com o objetivo de obter um padrão comparativo entre os algoritmos, visando a avaliação e escolha do modelo final, os testes realizados para o algoritmo Árvore de Decisão utilizaram a mesma metodologia citada anteriormente.

Além disso, para cada rodada de teste foram realizados testes do algoritmo de árvore de decisão visando evitar o overfitting[1], utilizando a técnica de "pré-poda", conforme orientado por Kumar (2021).

Os melhores resultados de acurácia média em modelos de árvore de decisão ocorreram em modelos no qual os dados passaram por padronização escalar, o valor de Random_State setado para em 101, Criterion com função "entropy" que prioriza o ganho de informação, alcançando respectivamente 65,69%, 77,88% e 73,76%. A figura 9 mostra as acurácias médias nos testes realizados.

1 Um cenário de overfitting ocorre quando, nos dados de treino, o modelo tem um desempenho excelente, porém quando utilizamos os dados de teste o resultado é ruim. Podemos entender que, neste caso, o modelo aprendeu tão bem as relações existentes no treino, que acabou apenas decorando o que deveria ser feito, e ao receber as informações das variáveis preditoras nos dados de teste, o modelo tenta aplicar as mesmas regras decoradas, porém com dados diferentes esta regra não tem validade, e o desempenho é afetado

Figura 9. Gráfico de acurácia média em testes de árvore de decisão com e sem padronização escalar



Fonte: Resultados da pesquisa

Os modelos gerados com o algoritmo de árvore de decisão com 10 validações por teste (tabela 11) obteve seu melhor resultado com 65,69% de precisão em teste sem padronização escalar, na terceira rodada de testes e com as seguintes alterações nos dados: inclusão dos dados de data no modelo, Criterion com a função "entropy" para o ganho de informação, Random_State definido para 101, Presort (divisão prévia de dados) ativado para o teste e Min_Samples_leaf com no mínimo duas amostras por folha.

Tabela 11. Resultados dos testes de Árvore de Decisão k=10.

Validações	Ajuste 1		Ajuste 2		Ajuste 3	
	Padronização Escalar		Padronização Escalar		Padronização Escalar	
	Sim	Não	Sim	Não	Sim	Não
Acurácia Média	**54,87**	**55,78**	**60,55**	**53,37**	**65,65**	**65,69**

Fonte: Elaborado pelo autor

A partir da rodada de testes 3 (tabela 12), em que houve inserção dos dados de data, foi possível perceber um ganho de acurácia nos resultados.

Tabela 12. Resultados dos testes de Árvore de Decisão k=30.

Validações	Ajuste 4		Ajuste 5		Ajuste 6	
	Padronização Escalar		Padronização Escalar		Padronização Escalar	
	Sim	Não	Sim	Não	Não	Sim
Acurácia Média	**77,88**	**64,74**	**61,82**	**67,86**	**68,4**	**72,07**

Fonte: Elaborado pelo autor

O melhor resultado dos testes com os dados para k igual a 30 foi de 77,88% de acurácia média (tabela 12) em teste realizado com padronização escalar, os parâmetros alterados para esta bateria de testes foram iguais aos parâmetros do melhor teste com k igual a 10. A diferença entre os dois testes foi a inclusão dos valores de data da ocorrência e não ocorrência e o numero de validações (k=30).

Tabela 13. Resultados dos testes de Árvore de Decisão k=29.

Validações	Ajuste 7		Ajuste 8	
	Padronização Escalar		Padronização Escalar	
	Não	Sim	Não	Sim
Acurácia Média	**72,87**	**73,76**	**72,07**	**64,78**

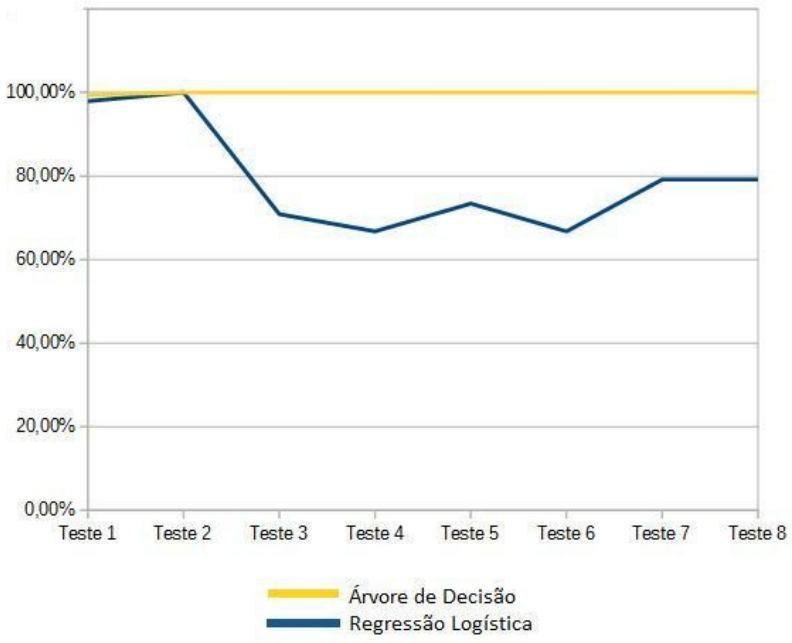
Fonte: Elaborado pelo autor.

Ainda com a finalidade de avaliar a sensibilidade do resultado ao número de validações, conforme testado também para os modelos de regressão logística, foi realizada uma nova bateria de testes para árvore de decisão, ou seja com k-29,. Nestes testes, o melhor modelo de árvore de decisão alcançou 73,76% de acurácia média (tabela 13), portanto, inferior ao resultado anterior, 77,88%, com k=30, em um teste realizado com padronização escalar de dados e parâmetros de configuração do modelo: Criterion com a função “entropy” para o ganho de informação, Random_State definido para 101, Presort (divisão prévia de dados) ativado para o teste.

5.4. Clusterização K-means

Para manter a imparcialidade dos testes realizados foram mantidos, nos testes de clusterização dos dados, parâmetros, valores de k, assim como o mesmo pré-processamento dos dados utilizado nos testes sem a aplicação da clusterização. A acurácia referente aos testes sem e com padronização escalar, com os mesmos parâmetros, foi semelhante, indicando que a aplicação desta função se faz desnecessária após a clusterização desta base de dados.

Figura 10. Acurácia média em testes utilizando o target da clusterização.



Fonte: Elaborada pelos autores.

O algoritmo de árvore de decisão, novamente apresentou resultados melhores que os modelos gerados a partir do algoritmo de regressão logística (figura 10). A acurácia média em testes daquele algoritmo variou na faixa de 99% e 100%, já os resultados para regressão logística variaram a acurácia entre 66,75% e 100%, demonstrando que para este algoritmo, os valores ótimos de Max_Iter e Tol definidos nos ajustes 1 e 2 (Max_Iter: 100 e 150; Tol:

0.0001) obtendo 97% e 100% de acurácia. Grandes alterações dos parâmetros citados reduziu significativamente a acurácia do algoritmo de regressão logística, levando a percepção que o algoritmo tem seu pico de resultado com menos iterações e menor tolerância ao critério de parada.

CONSIDERAÇÕES FINAIS

De uma forma geral, os resultados apresentados evidenciaram maior consistência nos resultados dos modelos de árvore decisão, o qual alcançou 77,88% de acurácia média, antes da implementação da clusterização, demonstrando uma consistência de resultados superior aos melhores resultados encontrados com regressão logística. Além disso, para este algoritmo, seus resultados foram elevados, em termos de acurácia, após a inclusão das datas de ocorrência e não ocorrência nos modelos concebidos, pois a diferença de 30 dias entre a NOC e OC ao inserir uma divisão a mais de nós da árvore de decisão, contribui para melhor caracterizar os grupos de dados, dado a elevada semelhança entre as variabilidades dos mesmos.

Os resultados permitiram evidenciar, ainda, que a clusterização beneficiou os modelos de árvore de decisão, que obteve resultados com acurácia persistente na faixa de 99 e 100% de acurácia (figura 10), resultados que confirmam a consistência do algoritmo para o problema de pesquisa estudado. Já o algoritmo de regressão logística apenas apresentou bons resultados com um conjunto específico de ajustes de parâmetros, pela alta variação em sua acurácia, podemos concluir que a clusterização elevou a acurácia da regressão logística, porém sem consistência em rodadas de testes com diferentes arranjos de parâmetros e diferentes quantidades de “k” pastas.

Ao compararmos com os resultados obtidos por Megeto (2012), 78% de acurácia no algoritmo de árvore de decisão, e com os melhores resultados obtidos pelos modelos sem clusterização, 77,88% e 54% de acurácia, na árvore de decisão e regressão logística, é possível concluir que o uso do algoritmo de clusterização, na etapa de preparação de dados, é fundamental para a elevação da acurácia. Podemos inferir que, conjunto de dados climáticos de pouca variação e correlação nula, como os dados deste trabalho, torna mais profícuo a capacidade de agrupamento a partir das distâncias euclidianas dos centróides de *clusters* realizados pelo algoritmo de clusterização k-means.

Os resultados permitem evidenciar que é possível produzir modelos preditivos a partir de dados abertos, utilizando o aprendizado de máquina e suas possibilidades, para

auxílio na gestão de produtores rurais no controle à ferrugem asiática, além disso, a elaboração de modelos preditivos para doenças, utilizando aprendizado de máquina, pode ser estratégica para uso comercial, principalmente considerando a técnica de clusterização, que permite maior acurácia para regiões com pouca variação climática, como no caso da região estudada.

Podemos citar como aplicação prática da metodologia empregada, um sistema de alerta com dados climáticos gerados a partir de previsões meteorológicas, apoiando o produtor rural na decisão do melhor momento para a aplicação de fungicidas. Em momento futuro se mostra interessante também validar a metodologia para os demais climas nacionais.

REFERÊNCIAS

AGRITEMPO (2021). **Sistema de Monitoramento Agrometeorológico**. Disponível em <https://www.agritempo.gov.br/agritempo/arquivos/Manual_Agritempo.pdf>. Acesso em 12/09/2021.

TOWARS DATA SCIENCE (2021). **03 Techniques to Avoid Overfitting of Decision Trees**. Disponível em : < https://towardsdatascience.com/3-techniques-to-avoid-overfitting-of-decision-trees-1e7d3d985a09>. Acesso em: 15/09/2021

ALMEIDA, A.M.R.; FERREIRA L.P.; YORINORI J.T.; SILVA J.F.V.; HENNING, A.A.; GODOY C.V.; COSTAMILAN L.M.; MEYER M.C. (2005). **Doenças da soja.** In: Kimati H. et al., (Eds.). Manual de Fitopatologia. São Paulo, Ceres. p.569-596.

AZEVEDO, E. **Desenvolvimento de Jogos e Aplicações em Realidade Virtual**. 3. ed., Rio de Janeiro: Elsevier, 2005.

BHOLOWALIA, Purnima; KUMAR, Arvind. **EBK-means: A clustering technique based on elbow method and k-means in WSN**. International Journal of Computer Applications, v. 105, n. 9, 2014.

BONDE, M. R.; NESTER, S. E.; AUTIN, C. N. et al. **Evaluation of virulence of Phakopsora pachyrhizi and P. meibomiae isolates**. **Plant Disease** 90(6):708-716.

2006.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovery Data**
Mining: From Concept to Implementation. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

CEPEA, 2021. **PIB do Agronegócio Brasileiro de 1996 a 2018. Participação do agronegócio no PIB do Brasil (em %)**. Disponível em:<https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx>. Acesso em: mai. 2021.

CEPEA: NÃO TRATAMENTO DE FERRUGEM NA SOJA PODE RESULTAR EM PERDA AO PRODUTOR DE QUASE R$ 12 BI. 2019. Disponível em: https://www.cepea.esalq.usp.br/br/releases/cepea-nao-tratamento-de-ferrugem-na-soja- pode-resultar-em-perda-ao-produtor-de-quase-r-12-bi.aspx. Acesso em: 15 set. 2021

CONSÓRCIO ANTIFERRUGEM (2021). **Embrapa Soja**. Disponível em <https://www.embrapa.br/soja/ferrugem/consorcioantiferrugem>. Acesso em 12/09/2021

COSTAMILAN, L. M.; P. F. BERTAGNOLLI; E J. T. YORINORI. 2002. **Perda de rendimento de grãos de soja causada por ferrugem asiática (Phakopsora pachyrhizi).** Fitopatologia Brasileira 27 (Suplemento): S100 (Resumo).

DUDA, Tanja; CANTY, Mort. Unsupervised classification of satellite imagery: choosing a good algorithm. **International Journal of Remote Sensing**, v. 23, n. 11, p. 2193-2212, 2002.

EMBRAPA. Disponível em: <https://www.infoteca.cnptia.embrapa.br/bitstream/doc/ 870114/1/COT2011161.pdf.>. Acesso em: mai. 2021.

FIGUEIRA, R. M. A. **Miner: um software de inferência de dependências funcionais**. Rio de Janeiro, 1998. Trabalho de Conclusão de Curso – Instituto de Matemática, Universidade Federal do Rio de Janeiro.

FONTANA, A., NALDI, M. C. **Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados**. 2009. Universidade de

São Paulo.

GALVANI FILHO, Mario Eugenio. **Detecção e quantificação de ferrugem asiática (Phakopsora pachyrhizi) na cultura da soja por imagens multiespectrais**. 2020. 60 f. Dissertação (Mestrado) - Curso de Proteção de Plantas, Instituto Federal Goiano, Urutaí, 2020.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4. ed. São Paulo: Atlas, 1987. 176 p.

GODOY, C.V. et al. **Eficiência de fungicidas para o controle da ferrugem-asiática da soja, Phakopsora pachyrhizi, na safra 2011/12: resultados sumarizados dos ensaios cooperativos.** Londrina: Embrapa Soja, 2012. 8p. (Embrapa Soja. Circular Técnica 93).

GODOY, C. V.; MEYER, M. C. Ensaio cooperativo para avaliação da eficiência de fungicidas no controle da ferrugem, Phakopsora pachyrhizi, em soja, na safra 2016/17, em londrina, PR. In: **Embrapa Soja-Artigo em anais de congresso (ALICE)**. In: REUNIÃO DE PESQUISA DE SOJA, 36., 2017, Londrina, PR. Resumos expandidos... Londrina: Embrapa Soja, 2017. p. 87-89., 2017.

HARTWIG, E. E. 1986. **Identification of a fourth major gene conferring resistance to soybean rust**. Crop Science 26: 1135-1136.

HARTMAN, G. L., J. B. Sinclair, and J. C. Rupe. 1999. **Compendium of Soybean Diseases.** ed. 4. APS Press, Minnesota.

HENNING, A. A.; GODOY, C. V. Situação da ferrugem da soja no Brasil e no mundo. In: ZAMBOLIM, L. (Ed.). **Ferrugem Asiática da Soja**. Viçosa: Universidade Federal de Viçosa, Departamento de Fitopatologia, p. 1-14. 2006.

IMEA. Instituto Mato-grossense de Economia Agropecuária. Disponível em:<http://www.imea.com.br/upload/publicacoes/arquivos/CPSoja_oUTUBRO_16.pdf>. Acesso em: set. 2021.

INMET (2018). Instituto Nacional de Meteorologia. Disponível em

<http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>. Acesso em 12/09/2021.

KOCHMAN, J. K. 1977. **Soybean rust in Australia**. In: Ford, R. E., and J. B. Sinclair. (Eds.). Rust of soybean – The problem and research needs. N° 12. University of Illinois, Urbana.

LIKAS, Aristidis; VLASSIS, Nikos; VERBEEK, Jakob J. **The global k-means clustering**

algorithm. Pattern recognition, v. 36, n. 2, p. 451-461, 2003.

MAY, Peter H. (Org.). **Economia do Meio Ambiente:** Teoria e Prática. 2. ed. Rio de Janeiro: Elsevier Editora Ltda., 2010.

MEGETO, G. A. S. **Avaliação da influência da temperatura e da precipitação na ocorrência da ferrugem asiática da soja por meio da técnica de árvore de decisão**. 2012. 94 f. Dissertação (Mestrado) - Curso de Engenharia Agrícola, Faculdade de Engenharia AgrÍcola, Universidade Estadual de Campinas, Campinas, 2012.

MICHALSKI, Ryszard Stanislaw; CARBONELL, Jaime Guillermo; MITCHELL, Tom M. (Ed.). **Machine learning: An artificial intelligence approach**. Springer Science & Business Media, 2013.

MILES, M. R., R. D. FREDERICK, AND G. L. HARTMAN. 2006. **Evaluation of soybean germplasm for resistance to Phakopsora pachyrhizi**. Plant Management Network, 2006. Consultado em 12 Julho de 2021. Publication in: <http://www.plantmanagementnetwork.org/pub/php/research/2006/germplasm>

Ministério da Agricultura, Pecuária e Abastecimento. **Mais de 67% das lavouras usam tecnologia na produção agrícola**. 2017. Acessado em: 26/06/2021. Disponível em: <http://www.agricultura.gov.br/noticias/tecnologia-ja-e-usada-em-cerca-de-67-das-propriedad es-rurais-do-pais-1>.

MOLIN, J. P.; AMARAL, L. R. do; COLAÇO, A. **Agricultura de precisão**. Em: Agricultura de precisão: resultados de um novo olhar. Brasília, DF, Embrapa, 2014.

NICOLINI, F.; REIS, E. M. ZOLDAN, E. M.; DANELLI, A. L. D.; ZANATTA, M.; AVOZANI, A.; TONIN, R. B.; GARCÉS, F. R. Efeito de temperaturas na germinação de uredosporos de *Phakopsora pachyrhizi* **Tropical Plant Pathology**, v. 35, n. 1,p. 143, 2010.

PAIVA, Roberto de Urzêda. **Algoritmo Paralelo para Processamento de Séries Temporais de Sensoriamento Remoto com Aplicação na Classificação do Uso e Cobertura do Solo**. 2021. 66 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal de Goiás, Goiânia, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. Cournapeau, D., Brucher,

M., Perrot, M. and Duchesnay, E.. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v12, p. 2825–2830

Portal Mais Soja (2021). Disponível em<http://maissoja.com.br/com-forte-presenca-daferrugem-nesta-safra-produtor-deve-estar-atent o-as-alternativas-de-controle-para-aproxima/>. Acesso em 22/09/2021.

PORTO, Marcelo Firpo; SOARES, Wagner Lopes. Modelo de desenvolvimento, agrotóxicos e saúde: um panorama da realidade agrícola brasileira e propostas para uma agenda de pesquisa inovadora. **Rev. Bras. Saúde Ocup.**, São Paulo, v. 01, n. 37, p. 17-50, maio 2012.

REIS, E. M.; CASA, R. T.; MICHEL, C. 2002. **Ocorrência de epidemia da ferrugem da soja no Rio Grande do Sul na safra 2001/2002.** Fitopatologia Brasileira 27 (Suplemento): S198 (Resumo)

REIS, Erlei Melo *et al*. **Mutações genéticas envolvidas com a resistência de Phakopsora pachyrhizi a fungicidas sítio-específicos e sua importância prática**. 2020. Disponível em: https://maissoja.com.br/mutacoes-geneticas-envolvidas-com-a-resistencia-de-phakopsora-pac hyrhizi-a-fungicidas-sitio-especificos-e-sua-importancia-pratica/. Acesso em: 01 ago. 2021.

RIBEIRO, R. **Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações.** Rio de Janeiro: IST-Rio, 2010.

RUSSELL, S.; NORVIG, P.. Inteligência Artificial. 3. ed. Rio de Janeiro: Elsevier, 2013.

SCHWALBERT, Raí Augusto. **Imagens de Satélite para Predição Espaço-Temporal da Produtividade de Milho e Soja em Diferentes Escalas Geográficas**. 2019. 94 f. Tese (Doutorado) - Curso de Engenharia Agrícola, Universidade Federal de Santa Maria, Santa Maria, Rs, 2019.

SINCLAIR, G. B.; BACKMAN, P. A. Compendium of Soybean Diseases (Translated by Rajabi, A.). University Publication Center, Tehran, 392 pp. **Persian with English summary**, 1993.

THAMADA, T. T. **Desenvolvimento de ensembles para alerta da ferrugem do cafeeiro em período crítico de progresso da doença.** 2016. 139 f. Dissertação (Mestrado) - Curso de Engenharia Agrícola, Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2016.

VAKILI, N. G.; BROMFIELD, K. R. 1976**. Phakopsora rust on soybean and other legumes in Puerto Rico**. Plant Disease 60:995-999.

VALE, Najla Kauara Alves do; ABDALA, Klaus de Oliveira. **MODELO TEÓRICO PARA A COMPREENSÃO DA INTERAÇÃO ENTRE AS VARIÁVEIS QUE INTERFEREM NA PRODUTIVIDADE DA SOJA**. Revista Eletrônica de Educação da Faculdade Araguaia, Goiânia, v. 3, n. 13, p.23-31, 21 mai. 2021.

Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace. **Portal do Agronegócio (2021)**. Disponível em<https://www.portaldoagronegocio.com.br/artigos/como-a-inteligencia-artificial-vai-transfom ar-o-agronegocio-3671>. Acesso em 22/09/2021.

VIEIRA, Simone Melo. **Detecção e classificação de soros lácteos em leites UHT e in natura empregando infravermelho médio com transformada de Fourier mais e análises multivariadas.** 2016.

WOLF, S. A.; WOOD, S. D. **Precision Farming: environmental legitimation, commodification of information, and industrial coordination. Rural Sociology, v. 62, n. 2, p. 180-206, 1997.**

YORINORI, J. T.; PAIVA, W. M. **Ferrugem da soja: Phakopsora pachyrhizi.** Londrina: Embrapa Soja, 2002. Folder

YORINORI, J. T.; PAIVA, W. M.; FREDERICK, R. D.; COSTAMILAN, L. M. et al. **Epidemics of soybean rust (Phakopsora pachyrhizi) in Brazil and Paraguay from 2001 to 2003**. Plant Disease, Saint Paul, v. 89, n. 6, p. 675-677, jun. 2005.