Ciência de dados aplicada a dados climáticos para inferir conhecimento associado ao desenvolvimento de Afídeos

Vinicius Rafael Schneider, Alexandre Tagliari Lazzaretti*

2021

Resumo

Os insetos possuem importante grau de colaboração para a manutenção do ecossistema no planeta. Porém ao atingir um determinado nível populacional e causar danos às plantas, alguns insetos, passam a ser considerados como insetos-pragas e representam uma ameaça para a agricultura. O afídeo ou pulgão, é um inseto que contém características para atingir este estado pois apresenta um alto potencial biótico e pode causar diferentes tipos de dano às plantas. Estudos comprovam que dados climáticos como precipitações, ventos e temperaturas interferem na quantidade populacional destes insetos. Portanto, este trabalho se propõe a aplicar técnicas e tecnologias de ciência de dados, a fim de verificar a correlação existente entre variáveis climáticas e a dinâmica populacional dos afídeos. Pode-se concluir que variáveis como precipitação, temperatura, quantidade de dias em que se chove na semana e fenômenos climáticos como El niño e La niña possuem influência na população de afídeos. Durante o trabalho, desenvolveu-se ainda quatro modelos, no intuito de predizer a população destes insetos. As acurácias dos modelo de predição desenvolvidos foram de 11,4% para Regressão Linear; 26,4% para o modelo de Rede Neural Artificial; 29,3% para Árvore de decisão e 41,4% para Floresta Aleatória.

Palavras-chaves: Extração de conhecimento, Análise exploratória, Regressão Linear, Redes Neurais Artificiais, Árvore de decisão, Floresta aleatória.

1 Introdução

A maioria das espécies de insetos é benéfica ou útil ao homem. Por executarem funções como polinização de plantas, decomposição de matéria orgânica, ter participação ativa no equilíbrio biológico, entre outros, os insetos acabam se tornando importantes na manutenção dos ecossistemas (FINKLER, 2011).

*vinirafaelsch@gmail.com, alexandrelazzaretti@ifsul.edu.br

Porém, segundo Gassen (1984) alguns insetos são considerados pragas quando alcançam níveis populacionais que possam causar danos às plantas, gerando assim redução no rendimento de grãos e consequentemente diminuição na lucratividade. Nesse sentido, Gassen (1984) cita ainda que, ao atingir o estado de praga, compensa ao agricultor o uso de métodos de controle para estes insetos.

O afídeo é um tipo de inseto que apresenta potencial para atingir o estado de praga quando alojado em determinada área. Wiest (2019) expõe a existência de aproximadamente 4.700 espécies classificadas na familia Aphididae, que é a família que agrupa os afídeos.

Em um estudo envolvendo diversas variáveis climáticas como precipitação, temperatura, ventos, Oliveira (1971) chegou à conclusão de que de uma forma ou de outra, com maior ou menor intensidade, os fatores climáticos tiveram o seu grau de atuação na diminuição ou aumento de pulgões.

Outro fator que possui correlação com a dinâmica populacional dos afídeos é o fenômeno ENOS (El Niño-Oscilação do Sul), pois confome citado em Cunha et al. (2001), a ocorrência deste fenômeno acaba alterando o padrão de circulação geral da atmosfera, o que acaba tendo influência no comportamento das variáveis climáticas.

Kamilaris, Kartakoullis e Prenafeta-Boldú (2017) aponta que a análise de dados possibilita à empresas e agricultores melhora na produtividade de suas lavouras através da extração de conhecimento ou detecção de padrões. A ciência de dados, segundo Aalst (2016), tem como objetivo gerar valor através de estudos voltados para uma grande quantidade de dados. Fazendo uso de disciplinas como mineração de dados, estatística, banco de dados, entre outras. Esta área torna possível realizar inferências, gerar gráficos e detectar padrões sobre um determinado assunto.

Sendo assim, seria interessante o uso de ciência de dados sobre base de dados coletados em lavouras e base de dados climáticos, no intuito de extrair conhecimento e gerar inferências baseando-se na correlação existente entre variáveis climáticas e dinâmica da população de afídeos, de forma a auxiliar entidades responsáveis na realização do controle populacional destes insetos.

Para isso, em um primeiro momento, foi organizada uma base de dados contendo dados climáticos para as regiões de monitoramento de afídeos. O que tornou possível verificar a correlação entre variáveis climáticas e a quantidade populacional dos afídeos. Foram avaliados os impactos causados pelos fenômenos El niño e La niña sob a população destes insetos e detectados períodos do ano em que ocorrem aumento ou diminuição na quantidade de afídeos coletados pelas armadilhas. Por fim, o desenvolvimento de modelos de predição para população de afídeo, com base nas variáveis e fenômenos climáticos abordados no trabalho.

O trabalho está estruturado da seguinte maneira: a seção 2 exibe o referencial bibliográfico. Na seção 3 são apresentados como se desenvolveu o trabalho e os resultados obtidos com ele. A seção 4 contém as considerações finais

2 Referencial bibliográfico

2.1 Afídeos

"Afídeos pertencem à família Aphididae, superfamília Aphidoidea, subordem Sternorrhyncha, ordem Hemiptera" (TOEBE, 2014, p. 23). Toebe (2014) expõe ainda que

este inseto possui formato ovalado, coloração variável e o tamanho máximo em que pode alcançar é de 5 milímetros de comprimento.

Segundo Stern (2008), os afídeos, também conhecidos como pulgões, possuem um corpo mole e pequeno. Se alimentam por meio da inserção de parte do seu intestino delgado na área responsável por conduzir o alimento para demais partes da planta, com o objetivo de usufruir dos nutrientes dispostos pela seiva da planta.

Wiest (2019) cita que os danos causados por estes insetos nas plantas podem ser classificados em dois tipos, direto e indireto. O dano direto acontece quando o afídeo se alimenta da seiva da planta e também através da inserção de toxinas presentes em sua saliva. O dano indireto acontece pela transmissão de vírus, como por exemplo o Vírus do Nanismo Amarelo da Cevada (VNAC), que conforme Gallo et al. (2002), adoece a planta, causando amarelamento em suas folhas e diminuição em seu tamanho.

"Plantas infestadas por esses homópteros apresentam folhas enroladas, encarquilhadas e raquíticas, sendo que a secreção açucarada excretada pelos mesmos reduz consideravelmente o valor comercial do produto" (AUAD; FREITA; BARBOSA, 2002, p. 335).

Os afídeos são classificados em três categorias durante sua vida: a fase inicial do afídeo intitula-se ninfa; ao entrar na fase adulta, podem variar entre duas classes, desta forma, são classificados como ápteros, que são os adultos que não possuem asas ou alado, adultos que possuem asas (STERN, 2008).

A dinâmica populacional destes insetos pode ser afetada por fatores bióticos causados por parasitóides, predadores e patógenos e também por fatores abióticos, que ocorrem através de variáveis como temperatura, umidade, luminosidade, entre outras. Sendo que a temperatura é um fator que apresenta interferência direta com o desenvolvimento populacional dos afídeos por afetar a taxa de desenvolvimento, reprodução e sobrevivência, refletindo assim, sua densidade populacional em uma planta (WIEST, 2019).

2.2 Dados climáticos

Segundo Torres et al. (2015), a medição de variáveis climáticas apresenta grande importância devido ao fato de que, diversas atividades humanas são direta e indiretamente afetadas por essas variáveis. Sendo assim, para suprir tal necessidade, criaram-se as estações meteorológicas de alta precisão e amostragem.

"Na agricultura, o monitoramento automático dos elementos meteorológicos tem contribuído não somente para o aumento da produtividade como, também, para a melhoria da qualidade dos produtos e para a preservação dos recursos naturais." (FERNANDES; ARAúJO; GOMES, 2013, p. 1)

Conforme visto no estudo realizado por Mafia et al. (2006), fatores como temperatura, precipitação pluviométrica e úmidade relativa do ar são exemplos de variáveis climáticas.

2.3 Enos (El Niño-Oscilação do Sul)

Segundo INPE... (2021), os fenômenos El niño e La niña fazem parte do fenômeno atmosférico-oceânico que ocorre no oceano Pacífico Equatorial, denominado El Niño Oscilação Sul (ENOS). Quando o oceano Pacífico Equatorial está mais quente que média

normal histórica, este evento é conhecido como El niño e quando o oceano Pacífico Equatorial está mais frio do que a média normal histórica, ocorre o La niña.

Se tratando da região sul do Brasil, em anos onde o fenômeno El niño ocorre existe um aumento na quantidade e ocorrência de precipitações, enquanto em anos de La niña, existe uma diminuição nos valores normais de chuva (CUNHA et al., 2001).

Cunha (1999) expõe que na safra de verão do sul do país, a ocorrência do fenômeno La niña apresenta condições favoráveis para o desenvolvimento de pragas de solo para o milho e desfavoráveis para o controle natural de pragas em soja. Rosenzweig et al. (2001) cita ainda que as condições climáticas durante os anos de El niño e La niña possuem correlação com danos causados por pragas em algumas regiões.

2.4 Banco de dados

Conforme Date (2004), o banco de dados por si só, pode ser considerado como o equivalente eletrônico de um armário de arquivamento, um recipiente para uma coleção de arquivos de dados. Enquanto o SGBD é uma ferramenta que torna possível a manipulação desses dados armazenados no banco de dados.

Portanto, um banco de dados que é gerenciado por um SGBD, torna-se um sistema que permite gerenciar informações em forma de registros. Através deste sistema é possível resgatar, alterar e excluir registros conforme demanda de usuários.

Um exemplo de banco de dados é o AgroDB, que será utilizado para este trabalho. Segundo Lazzaretti et al. (2016a), o AgroDB é um banco de dados que serve como repositório de dados agronômicos e possui modelagem genérica, que facilita a integração com modelos de simulação de safra.

2.5 Matriz de correlação

A Matriz de Correlação é utilizada para iniciar a análise estatística dos dados, através dela é possível identificar, de forma visual, as variáveis estudadas que se relacionam entre si (LIDIANE et al., 2018). O cálculo utilizado para matriz de correlação é o Coeficiente de correlação de *Spearman*, que segundo Sousa (2019, p. 19), "avalia a intensidade e o sentido da relação monótona entre duas variáveis".

Este cálculo resulta em um valor entre -1 e 1 responsável por expressar o grau de dependência entre duas variáveis. Quando negativas, significa que uma variável diminui com o aumento da outra, caso positiva, significa que uma variável aumenta com o aumento da outra (SOUSA, 2019).

Segundo Solagna (2020), quando o resultado obtido no coeficiente de correlação for de 0,1 até 0,3, significa que a correlação entre as variáveis é fraca. De 0,4 até 0,6, a correlação é moderada ou média. Para 0,7 até 1, tem-se uma correlação forte. Valores abaixo de 0,1 não são considerados.

2.6 Regressão Linear

Segundo Rodrigues, Medeiros e Gomes (2013), a análise de regressão busca analisar dados com o objetivo de saber se, e como duas ou mais variáveis relacionam-se. Esta análise tem como resultado uma equação matemática que descreve o relacionamento entre essas variáveis.

Rodrigues, Medeiros e Gomes (2013) expõe ainda que esta análise "pode ser usada para estimar ou predizer, valores futuros de uma variável quando se conhecem ou se supõem conhecidos valores da outra variável".

A equação linear é definida pelas seguintes variáveis: α e β, que correspondem respectivamente, ao coeficiente linear da reta e ao coeficiente angular, que são parâmetros do modelo. A variável a ser predita corresponde a γ e $\times$ é a variável independente, ou seja, a variável em que se acredita alterar o comportamento da variável γ. Por fim, a variável ϵ, que corresponde às variáveis aleatórias que correspondem ao erro, explicando a variância existente em γ, que não consegue ser explicada por $\times$ (RODRIGUES; MEDEIROS; GOMES, 2013).

$$\gamma = \alpha + \beta \times i + \epsilon i \tag{1}$$

2.7 Redes Neurais Artificiais

Conforme VENDRUSCOLO et al. (2015), é um sistema computacional paralelo constituído de uma ou mais unidades simples de processamento, também conhecidos como neurônios, que conectados entre si são encarregados de calcular determinadas funções matemáticas que, normalmente, são não-lineares.

VENDRUSCOLO et al. (2015) cita ainda que os neurônios são dispostos em uma ou mais camadas interligadas por um grande número de conexões, geralmente unidirecionais. Na maioria dos modelos as conexões são associadas a pesos, que armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede.

Na Rede Neural Artificial, a camada de entrada é responsável somente por receber os valores de entrada, enquanto as camadas intermediárias e a camada de saída são responsáveis por mapear o conhecimento e através de seus neurônios processar as informações, realizando uma série de cálculos com base em pesos armazenados nas conexões e valores contidos em cada neurônio. Um resultado é gerado através da última camada da rede, podendo ser uma classificação ou estimativa (VENDRUSCOLO et al., 2015).

2.8 Árvore de decisão

A árvore de decisão pode ser utilizada como classificador ou preditor e possui duas etapas: a de aprendizado e a de classificação/predição. Na fase de aprendizado é gerada uma árvore de decisão com base no conjunto de amostras utilizado para o treinamento. Já a etapa de classificação ou regressão é quando a árvore faz uso da estrutura gerada na fase de aprendizado para classificar ou predizer um determinado dado (LAN et al., 2020).

Segundo Lan et al. (2020), sua estrutura é parecida com a de um fluxograma, que pode ser dividido em três partes: os nós internos, as ramificações e o nós folhas. Cada um dos nós internos representam um teste gerado na fase de aprendizado da árvore, as ramificações correspondem a uma saída dos nós internos, gerando assim um caminho para um nó folha, que por sua vez representa a saída de dados, ou seja, uma classificação ou predição.

2.9 Floresta Aleatória

A floresta aleatória aproveita a técnica de árvore de decisões para realizar classificação de dados. Lan et al. (2020) cita que a floresta aleatória combina várias árvores de

decisões, intutuladas como classificadores fracos, para assim formar um classificador mais forte. Cada árvore utiliza um determinado número de amostras aleatórias para realizar o treinamento, garantindo assim a aleatoriedade dos dados.

Para gerar a classificação ou predição dos dados, a floresta aleatória faz com que as árvores de decisão elejam, através de votação, a classe mais apropriada para uma entrada de dados (LAN et al., 2020).

2.10 Python

Python é uma linguagem de programação de alto nível, orientada a objetos, com sintaxe concisa e que favorece a legibilidade do código-fonte. A linguagem possui tipagem forte e dinâmica, interpretada e interativa. O *Python* conta ainda com estruturas de alto nível, como listas, dicionários, data/hora, entre outros, além de módulos prontos, *frameworks* criados e disponibilizados por terceiros. Possui recursos como geradores, introspecção, persistência e unidades de testes (BORGES, 2014).

Esta linguagem de programação foi escolhida para o desenvolvimento do trabalho pois, conforme Pedregosa et al. (2011), a linguagem está se estabelecendo como uma das linguagens mais populares para computação científica por apresentar natureza interativa de alto nível e um amadurecimento em seu ecossistema de bibliotecas científicas, facilitando o desenvolvimento algorítmico e a análise exploratória de dados.

2.11 Scikit-learn

O *Scikit-learn* ou *Sklearn*, é uma biblioteca de aprendizado de máquina de código aberto projetada e escrita em *Python*. Abraham et al. (2014) cita que a biblioteca fornece a implementação de eficientes algoritmos de última geração para aprendizado de máquina, de forma que até mesmo especialistas de outras áreas consigam fazer uso da mesma.

No *scikit-learn*, todos os objetos e algoritmos aceitam dados de entrada de matrizes bidimensionais na forma de registros e colunas. Os objetos *Scikit-learn* compartilham um conjunto uniforme de métodos que depende de sua finalidade: estimadores podem ajustar modelos de dados, preditores podem fazer previsões sobre novos dados e transformadores convertem dados de uma representação para outra. (ABRAHAM et al., 2014, p. 2)

Conforme Pedregosa et al. (2011), a biblioteca inclui algoritmos de aprendizado de máquina, sendo eles do tipo supervisionado e não supervisionado. Alguns dos algoritmos presentes na biblioteca são: *k-means*, que faz uso da metodologia de *clusters*, e classifica objetos de acordo com a proximidade de suas características com as características de determinado *cluster*, floresta aleatória que cria uma árvore aleatória, onde cada nó representa uma característica de um objeto e os nós folhas representam um objeto cadastrado na base de dados, ou seja, uma saída/classificação.

Para calcular a acuráciados modelos desenvolvidos no trabalho utiliza-se o coeficiente de determinação R^2 sobre os dados separados para teste. Conforme documentação presente em Scikit-learn... (2021), o coeficiente de determinação é calculado através presente equação, onde u é a soma residual dos quadrados e v é a soma total dos quadrados. A melhor pontuação possível para o coeficiente de determinação R^2 é 1,0 que, transformando para porcentagem pode ser visto como 100%.

$$(1 - \frac{u}{v}) \tag{2}$$

2.12 Pandas

Pandas é uma biblioteca *Python* voltada para manipulação e análise de dados. McKinney (2019, p. 4) explica que "a biblioteca fornece estruturas e funções de dados ricas, projetadas para tornar o trabalho com dados estruturados rápido, fácil e expressivo".

Segundo McKinney (2011, p. 1), "a biblioteca fornece rotinas intuitivas integradas para a execução de manipulações e análises de dados comuns em tais conjuntos de dados". Através dela é possível indexar, separar, selecionar e agrupar subconjuntos de dados para tornar mais fácil a manipulação e análise dos mesmos (MCKINNEY, 2019).

Possui como principal objeto o *dataframe*, que implementa diversas funcionalidades presentes na linguagem de análise de dados R. O *dataframe* é uma estrutura de dados tabular bidimensional, formada por linhas que são indexadas e colunas, podendo ser visto como uma estrutura de dicionário (MCKINNEY, 2019).

3 Resultados

3.1 Estudo de caso

Para o desenvolvimento deste trabalho, foi utilizada a base de dados proveniente do *Trapsystem*, que segundo Lazzaretti et al. (2016b), é uma aplicação que "permite o gerenciamento de dados obtidos pela captura/coleta de insetos em armadilhas". Os dados desta aplicação foram cedidos pela Embrapa (Empresa Brasileira de Pesquisa Agropecuária) e possui a quantidade de afídeos coletados em lavouras. A espécie destes pulgões que foi disponibilizada e será usada no estudo é a *Rhopalosiphum padi* (Linnaeus, 1758) que segundo Wiest (2019) é uma praga que apresenta ampla distribuição geográfica e abundância.

A base de dados de afídeos foi exportada em formato *csv* e, conforme figura 1, têm seus dados armazenados de forma semanal possuindo os seguintes campos: *station*, que armazena um identificador para uma determinada estação; *date*, referente a data em que o dado foi coletado; semana, semana do ano em que o dado foi coletado; *id*, refere-se a uma identificação de armadilha; *description*, campo para armazenar um nome para uma armadilha; *desinsect*, uma descrição da operação feita para chegar ao resultado da coluna *total*; por fim, *total*, responsável por armazenar a quantidade de afídeos encontrados na ocasião.

```
station; date; semana; id; description; desinsect; total
435;2017-01-05;1;3;Armadilha 0;Soma Agrupada de Afideos;4
435;2017-01-12;2;3;Armadilha 0;Soma Agrupada de Afideos;2
435;2017-01-19;3;3;Armadilha 0;Soma Agrupada de Afideos;2
435;2017-01-26;4;3;Armadilha 0;Soma Agrupada de Afideos;14
435;2017-02-02;5;3;Armadilha 0;Soma Agrupada de Afideos;25
435;2017-02-09;6;3;Armadilha 0;Soma Agrupada de Afideos;30
435;2017-02-16;7;3;Armadilha 0;Soma Agrupada de Afideos;14
435;2017-02-23;8;3;Armadilha 0;Soma Agrupada de Afideos;44
435;2017-03-02;9;3;Armadilha 0;Soma Agrupada de Afideos;19
```

Figura 1 – Dados de afídeos coletados em armadilhas.

Outra base de dados utilizada no estudo é referente a dados climáticos que são armazenados e disponibilizados de forma pública pelo INMET (Instituto Nacional de Meteorologia).

A coleta dos dados climáticos é feita de forma diária e seus dados são disponibilizados em formato *json*, possuindo, conforme figura 2, os campos para armazenamento de: *id*, que armazena um identificador para os registros que ocorrem; idEstacao: campo para

identificação da estação; data, armazena a data em que o dado é coletado; temperaturaMinima, temperaturaMedia e temperaturaMaxima, que representam, respectivamente, a temperatura mínima, média e máxima que aconteceu em um determinado dia; precipitacao, quantidade em milímetros de chuva que aconteceu em um dia.


```
[
  {
    "id": 127153849,
    "idEstacao": 9003000,
    "data": "2017-01-01",
    "temperaturaMinima": 18.08,
    "temperaturaMinimaDuvidosa": false,
    "temperaturaMinimaEstimada": true,
    "temperaturaMedia": 23.1,
    "temperaturaMediaDuvidosa": false,
    "temperaturaMediaEstimada": true,
    "temperaturaMaxima": 28.12,
    "temperaturaMaximaDuvidosa": false,
    "temperaturaMaximaEstimada": true,
    "precipitacao": 29.1,
    "precipitacaoDuvidosa": false,
    "precipitacaoEstimada": true,
    "editavel": false
  },
```


Figura 2 – Dados climáticos do INMET.

As cidades e os períodos levados em consideração para realização deste trabalho podem ser vistos na tabela 1.

	Período	
Cidade	**Dados de afídeos**	**Dados climáticos**
Coxilha	jan/2017 à dez/2020	jul/2018 à dez/2020
Passo Fundo	jan/2017 à dez/2020	jan/2017 à dez/2020

Tabela 1 – Cidades e períodos que estão sendo estudados.

Para verificar a correlação existente entre a ocorrência do fenômeno ENOS com a dinâmica populacional dos afídeos se fez necessária a descoberta dos períodos em que estes fenômenos aconteceram. A aquisição destes dados se deu através do site da NOAA (*National Oceanic and Atmospheric Administration*), onde a temperatura do oceano é medida de forma mensal.

A fim de determinar que tipo de fenômeno aconteceu em um determinado ano foi levado em consideração a ocorrência dos fenômenos entre os períodos de agosto de um ano até julho do ano seguinte, assim como é feito e exposto no site do INPE (Instituto Nacional de Pesquisas Espaciais). A tabela 2 expõe a ocorrência de fenômenos ENOS em seus respectivos períodos.

Período	**Fenômeno ENOS**
ago/2016 à jul/2017	La Niña
ago/2017 à jul/2018	La Niña
ago/2018 à jul/2019	El Niño
ago/2019 à jul/2020	Ano Neutro
ago/2020 à jul/2021	La Niña

Tabela 2 – Ocorrência de fenômenos ENOS e seus períodos.

3.2 Metodologia

Conforme figura 3, utilizando ciência de dados (passo 1) sobre as bases de dados (passo 2) a etapa de seleção de dados (passo 4) foi realizada, nela ocorreu a leitura dos dados climáticos e dados de afídeos.

Na fase de pré-processamento (passo 5) ocorre a manipulação dos dados, nela são analisadas e tratadas possíveis falhas ou inconsistências. Foi desenvolvido um *dataframe* contendo dados de afídeos e dados climáticos para cada cidade e também um *dataframe* que concatena os dados correspondentes à todas as cidades estudadas. Também foram adicionadas duas novas colunas referentes a ocorrência do fenômeno ENOS e o número de dias em que se choveu na semana para cada cidade.

A etapa de *Data Mining* (passo 6) é de exploração dos dados, onde ocorre a extração de conhecimento. Para isso foi feita a plotagem e análise de gráficos, utilização da matriz de correlação e o uso de técnicas de aprendizado de máquina com o objetivo de predizer a quantidade de afídeos com base nas variáveis climáticas abordadas.

Ainda se tratando do passo 6, na análise exploratória as cidades estudadas foram analisadas de forma individual, isso foi feito para verificar se a correlação entre variáveis climáticas e quantidade de afídeos se faz verdadeira e não seja algo ocasional. Para as demais etapas, foi feita a concatenação dos dados, unindo dados referentes a todas as cidades em um único dataframe.

Por fim, a última etapa é a de avaliação de resultados (passo 7), onde foi avaliada a existência de correlação entre dados climáticos e dados de afídeos e os resultados obtidos com modelos de predição desenvolvidos.

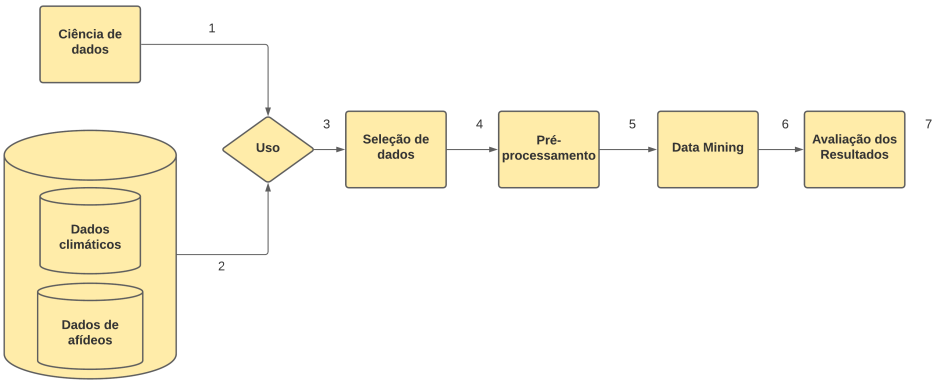


Figura 3 – Metodologia do trabalho.

3.3 Seleção de dados

A leitura dos arquivos contendo os dados foi feita utilizando funções disponibilizadas pela biblioteca *Pandas*, que permite executar a leitura de arquivos com extensões *csv* e *json*, transformando-os em um novo *dataframe*.

Sendo assim, obteve-se como resultado duas novas estruturas de *dataframe* para cada cidade, sendo que uma delas é referente aos dados de afídeos e outra tratando-se dos dados climáticos.

3.4 Pré-processamento

Na etapa de pré-processamento foram removidas as colunas julgadas desnecessárias para o estudo dos dados e o treinamento dos modelos. Nos dados de afídeos foram eliminados os campos: id, estacao, descricao e *desinsect*. Para os dados climáticos, julgou-se necessária a permanência de todas as variáveis.

Como os dados de afídeos são armazenados de forma semanal, optou-se por transformar o *dataframe* de dados climáticos em dados semanais. Para isso, foi feita uma média semanal dos campos temperaturaMinima, temperaturaMedia e temperaturaMaxima. No campo precipitacao, foi feita a soma do total de chuva que ocorreu na semana.

Para tornar possível as operações de média e soma, o primeiro passo foi gerar duas novas colunas no *dataframe* dos dados climáticos, uma contendo a semana do ano e outra contendo o ano em que o registro ocorreu. Isso foi feito utilizando a *Pandas*, que possibilita extrair essas informações a partir do campo que armazena a data de cada registro.

Por meio das colunas criadas e agrupamento de dados, foi possível executar operações de média e soma sobre os dados. A partir disto, foi criado um novo *dataframe*, que armazena os dados semanais com os campos de: ano, semana, temperaturaMinima, temperaturaMedia, temperaturaMaxima e precipitacao.

Pelo fato de algumas cidades conter mais de uma armadilha, foi feita a média de ocorrência em cada semana para cada cidade, posterior a isso, estes dados foram concatenados ao novo *dataframe* utilizando laços de repetição, onde os mesmos são atribuídos a uma determinada linha caso a semana e o ano forem iguais nos dois *dataframes*. Isso foi feito pois através de observações realizadas sobre a base de afídeos constatou-se a existência semanas em que nenhum dado é cadastrado. Após este passo, foram eliminados os registros cujo campo de afídeos era *NaN*, significando semanas em que a ocorrência de afídeos não foi cadastrada na base de dados de afídeos.

A fim de verificar a influência que os fenômenos climáticos ENOS têm sobre a dinâmica populacional dos afídeos e também com objetivo de melhorar a acurácia dos modelos treinados, foi adicionada uma coluna chamada fenomeno, que armazena o fenômeno climático que aconteceu em um determinado ano. Nesta nova coluna, o valor 0 corresponde à ocorrência do El niño, 1 para La niña e 2 para anos neutros, que significa que nenhum dos dois fenômenos aconteceram.

No intuito de extrair o máximo possível dos dados estudados e também, tendo como base o que disse Cunha et al. (2001), onde argumenta que não só a quantidade mas também a ocorrência de chuva é maior em anos de El niño, foi adicionado em cada *dataframe* um campo chamado frqChuva, que armazena a quantidade de dias em que se choveu na semana. Para a contabilização de um novo dia de chuva na semana foi levado em consideração qualquer tipo de valor maior que 0 para a coluna de precipitacao.

Portanto, como forma de exemplificação a figura 4 representa o *dataframe* final referente à cidade de Passo Fundo. Vale ressaltar que o mesmo processo foi realizado para as demais cidades estudadas. Também nesta etapa ocorreu a criação de um *dataframe* que concatena os dados de todas as cidades, que será utilizado posteriormente na matriz de correlação e treinamento de modelos de predição.

	ano	semana	temperaturaMinima	temperaturaMedia	temperaturaMaxima	precipitacao	frqChuva	fenomeno	afideos
0	2017	1	19.63	24.23	28.82	39.8	6.0	1.0	4.0
1	2017	2	18.13	23.35	28.56	23.2	4.0	1.0	2.0
2	2017	3	17.89	23.24	28.58	16.3	2.0	1.0	2.0
3	2017	4	15.50	21.10	26.70	19.0	2.0	1.0	14.0
4	2017	5	17.96	22.41	26.86	22.9	7.0	1.0	25.0
...	...	...	...	...	...	...	...	...	...
173	2020	47	12.55	19.50	26.45	6.3	5.0	1.0	34.0
174	2020	48	17.47	23.37	29.28	68.8	4.0	1.0	10.0
175	2020	49	16.25	21.21	26.17	51.1	5.0	1.0	5.0
176	2020	50	16.73	23.70	30.67	2.0	2.0	1.0	4.0
177	2020	51	18.29	23.14	27.98	13.3	3.0	1.0	11.0

Figura 4 – Dataframe final da cidade de Passo Fundo.

3.5 Data Mining

Conforme citado na metodologia, nesta etapa é feita a extração de conhecimento e geração de inferência sobre os dados estudados.

Com o objetivo de extrair conhecimento, em um primeiro momento foi realizada a análise exploratória sobre os dados no intuito de verificar a correlação existente entre dados climáticos com a dinâmica populacional dos afídeos. Para isso, foram realizadas plotagens e análise de gráficos e também a utilização da matriz de correlação aplicando o coeficiente de correlação de *Spearman*.

Após isso, foram aplicadas técnicas de aprendizado de máquina sobre os dados no intuito de predizer a população de afídeos a partir de dados climáticos. As técnicas utilizadas neste processo foram Regressão Linear, Redes Neurais Artificiais, Árvore de decisão e Floresta Aleatória. A escolha das técnicas de predição foram feitas principalmente com base na leitura de artigos científicos.

Optou-se por implementar Regressão Linear pois Rodrigues, Medeiros e Gomes (2013) investigou o uso desta técnica para realização de inferências e concluiu sua viabilidade para predição de dados. Nascimento, Junior e Fagundes (2018) cita ainda que o uso desta técnica pode ser feito quando predição exata dos valores não é necessária e sim, uma aproximação, o que se encaixa no objetivo do trabalho.

Foi detectada a não linearidade no comportamento populacional dos afídeos com base na análise exploratória que será exposta em seguida. Sendo assim, a técnica de Redes Neurais Artificiais foi testada com base no que foi exposto no referencial bibliográfico, onde é citado que o modelo de saída gerado a partir do treinamento de uma Rede Neural Artificial não é linear.

A Árvore de decisão foi implementada pois Dias (2002) classifica esta técnica como adequada para realizar predições de dados contendo variáveis numéricas. Neste sentido, através de pesquisas feitas na documentação da biblioteca Scikit-learn, julgou-se interessante a implementação da Floresta Aleatória pois a mesma faz uso de diversas Árvores de decisão, gerando como saída o dado com maior votação entre as árvores.

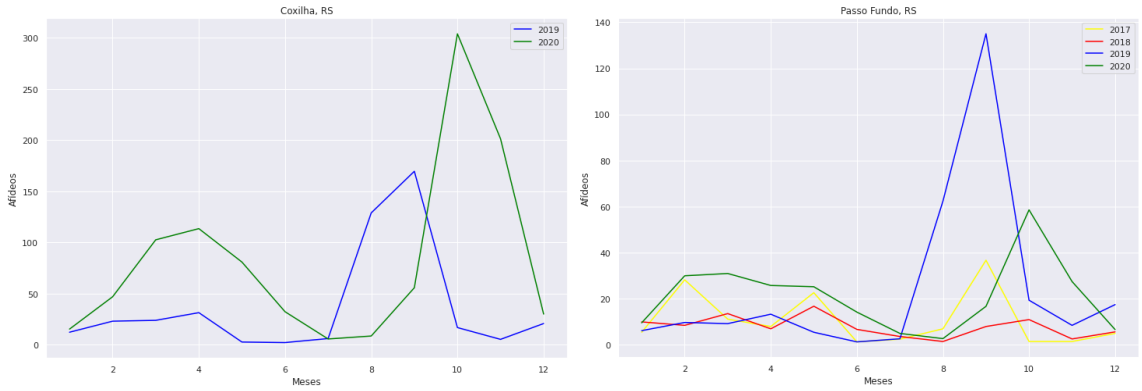
No intuito de melhorar a assertividade dos modelos desenvolvidos foi utilizado um recurso chamado *GridSearchCV*. Este recurso permite que diferentes configurações de treinamento sejam passadas como parâmetro, a partir disso são feitas todas combinações possíveis, retornando a que obteve melhor acurácia. O recurso *GridSearchCV* e as técnicas de predição supracitadas estão presentes na biblioteca *Scikit-learn* e suas aplicações foram desenvolvidas com base em estudos voltados sobre a documentação da biblioteca.

Para realizar o treinamento dos modelos de predição foram utilizados dados referente à todas cidades, sendo que 80% dos dados foram separados de forma aleatória para a etapa de treinamento e os 20% restantes foram utilizados para comparações e testar a acurácia dos modelos. As variáveis climáticas utilizadas para o treinamento de ambos modelos foram precipitacao, semana, temperaturaMinima, temperaturaMaxima, frqChuva e fenomeno.

3.5.1 Análise exploratória

O primeiro passo dado na etapa de análise exploratória foi avaliar o comportamento da população de afídeos durante o período estudado. Para isso foi feita uma análise mensal sobre os dados referente à cidade de Coxilha no período de jan/2019 à dez/2020 e Passo Fundo no período de jan/2017 à dez/2020. Estes períodos foram escolhidos por apresentarem maior frequência no armazenamento dos dados de afídeos.

Na figura 5 é exposto a dinâmica populacional de afídeos da cidade de Coxilha (A) e Passo Fundo (B). Nela, cada ano estudado é representado por uma linha no gráfico, sendo que, o eixo X representa o meses de cada ano e o eixo Y a média de afídeos encotrados em cada mês. A análise de ambos os gráficos permite perceber que a ocorrência de afídeos não ocorre de forma linear.



(A) (B)

Figura 5 – Comportamento mensal da população de afídeos.

Foi observado ainda que nos meses de junho, julho e agosto costuma ocorrer uma diminuição na quantidade populacional dos afídeos, este período corresponde a estação de inverno. Outro padrão que se encontrou foi que posterior a este período, nos meses de setembro e outubro, na maioria das ocasiões aconteceram picos na ocorrência destes pulgões, os meses fazem parte do final de inverno e se estende por grande parte da primavera.

Os padrões encontrados entre as épocas do ano e a população de afídeos estão de acordo com estudos feitos por PEREIRA, Lau e JÚNIOR (2016), onde detectou que reduções significativas nos pulgões acontecem nos meses de junho, julho e agosto. Neste mesmo artigo é citado ainda os picos populacionais observados entre os meses de setembro e outubro. Tomé et al. (2013) corrobora expondo que os picos que acontecem nas estações de inverno e primavera representam de 89 a 95% das ocorrências de uma espécie de afídeos conhecida como *Rhopalosiphum padi* (Linnaeus, 1758) na localidade de Coxilha, RS.

Optou-se por verificar a correlação existente entre variáveis climáticas e a dinâmica populacional dos afídeos pois Oliveira (1971) conclui em seu estudo que fatores como temperatura e precipitações possuem correlação com a quantidade de afídeos. Rosenzweig

et al. (2001) argumenta ainda que anos de fenômenos El niño e La niña também possuem correlação com danos causados por pragas.

Sendo assim, foi realizada a análise de gráficos de forma semanal sobre cada uma das cidades em seus respectivos períodos expostos na tabela 1. Esta análise pode ser observada na figura 6, onde cada linha de gráficos representa o estudo feito sobre determinada cidade. Nela, cada ponto presente nos gráficos representa o ponto de intersecção entre o eixo Y, que representa a ocorrência de afídeos na semana com o eixo X, que representa em diferentes gráficos as variáveis de temperaturaMedia, precipitacao, frqChuva e fenomeno.

Com base na análise dos gráficos da figura 6, notou-se que a população de afídeos diminui em semanas de temperaturas extremas e apresenta maior quantidade em semanas onde a temperatura média fica entre 15 e 22°C. O mesmo padrão foi exposto em PEREIRA, Lau e JÚNIOR (2016) e Tomé et al. (2013), onde a ocorrência de populações de afídeos acima da média foram detectadas em períodos em que a temperatura média se manteve entre 15 e 20°C. Esses mesmos estudos citam ainda picos na população de afídeos para semanas em que a precipitação se manteve abaixo de 20 mm para PEREIRA, Lau e JÚNIOR (2016) e 30 mm para Tomé et al. (2013), conforme os gráficos expostos, os mesmos padrões podem ser observados.

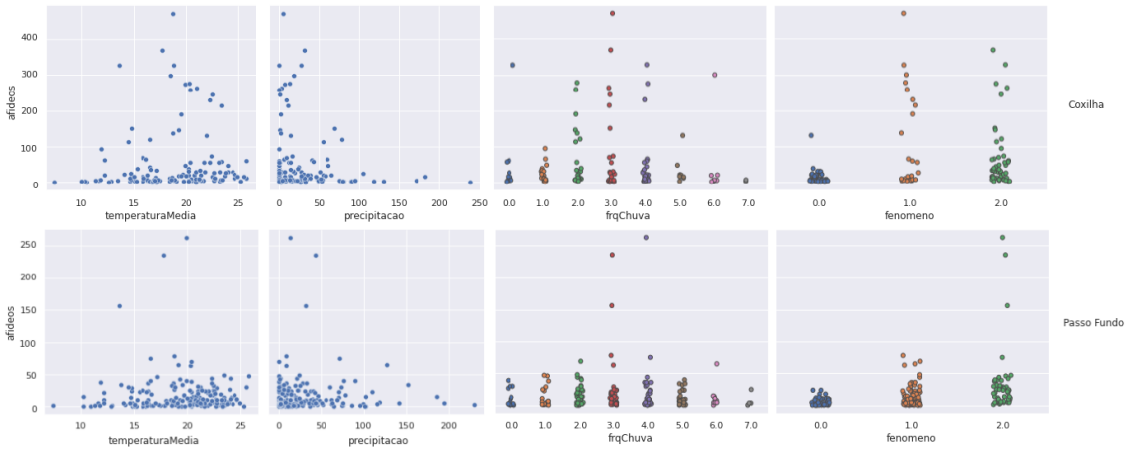


Figura 6 – Gráficos relacionando dados climáticos e ocorrência de afídeos para cada cidade.

Ainda se tratando da figura 6, através da análise do campo frqChuva este estudo supõe que a ocorrência de afídeos acima da média possui ambiente mais propício para acontecer quando chove-se entre 2 à 4 dias durante a semana. Quanto aos fenômenos ENOS, o gráfico expõe diminuição considerável na população dos pulgões em anos de El niño (0) quando comparados a anos de La niña (1) e anos neutros (2). Tal fato confirma a correlação exposta por Rosenzweig et al. (2001) entre a dinâmica populacional de afídeos com anos de fenômeno El niño e La niña.

Foi feito ainda o uso da matriz de correlação para verificar os níveis de correlação existentes entre as variáveis através de cálculos matemáticos. O cálculo de correlação de *Spearman* foi utilizado na matriz de correlação pois conforme Sousa (2019) pode ser utilizado para avaliação de dados não lineares. Conforme a figura 6, as duas cidades apresentam o mesmo padrão na população de afídeos, sendo assim, optou-se por concatenar os dados das cidades estudadas para então aplicar a matriz de correlação.

Para fazer a análise da matriz presente na figura 7 basta associar a linha da variável

desejada com a coluna de outra variável desejada, ou vice-versa. Sendo assim, o cálculo de *Spearman* mostra que a quantidade populacional de afídeos possui correlação com as variáveis temperaturaMinima (0,13), temperaturaMedia (0,15), temperaturaMaxima (0,17), precipitacao (0,1) e fenomeno (0,37).

O fato das correlações entre temperaturas e afídeos serem positivas mostra que grandes populações de afídeos tendem a ter dificuldades em ocorrer em temperaturas partindo de 0º, conforme a temperatura aumenta esta probabilidade vai crescendo também. O mesmo ocorre com a variável de precipitação, que por se tratar de uma correlação negativa, quer dizer que quanto mais milímetros de chuvas ocorrem por semana, menores são as populações de afídeos.

Segundo o cálculo de *Spearman*, a variável contendo fenômenos ENOS é a que apresenta maior correlação com a população destes insetos (0,37). Acredita-se que o motivo desta variável apresentar uma correlação consideravelmente maior em relação as outras é que a ocorrência destes fenômenos acaba alterando diversas variáveis climaticas não estudadas, como por exemplo umidade e os ventos que conforme Galdino et al. (2007), são variáveis não estudadas que alteram a quantidade populacional dos afídeos.

Vale ressaltar que o fato de que algumas variáveis não apresentam nenhum tipo de correlação na matriz não significa que as duas variáveis não tenham correlação entre sí, pois, conforme explica Sousa (2019), a não existência de correlação entre duas variáveis, não significa que não se verifique outro tipo de correlação diferente da que foi aplicada, como por exemplo, a correlação linear ou exponencial.

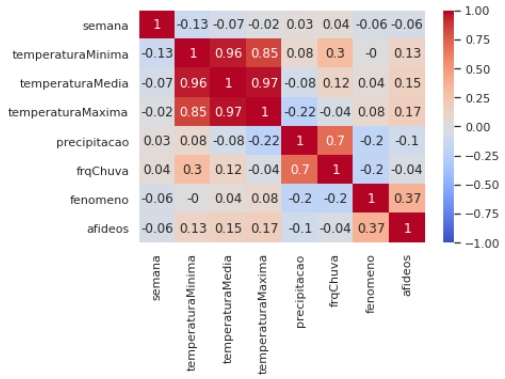


Figura 7 – Matriz de correlação aplicada à dataframe contendo dados de todas cidades.

3.5.2 Regressão Linear

O desenvolvimento da técnica de Regressão Linear se deu por meio da classe *Linear-Regression* presente na biblioteca do *Scikit-learn*. Conforme a documentação da biblioteca, a técnica de regressão linear não apresenta muitas configurações para o treinamento do modelo. Portanto o único passo executado para esta técnica foi o de treinamento sobre a base de dados que foi separada para etapa de treinamento dos modelos.

Para visualizar as predições realizadas pelo modelo de regressão linear, a figura 8 mostra as ocorrências da base de dados de testes, cada ocorrência possui o valor real de afídeos (linha vermelha) e o valor predito pelo modelo (linha azul). O eixo X representa as ocorrências e o eixo Y a quantidade de afídeos. Utilizando o coeficiente de determinação

R^2 para verificar a acurácia do modelo de regressão linear sobre os dados separados para testes chegou-se a uma acurácia de 11,4%.

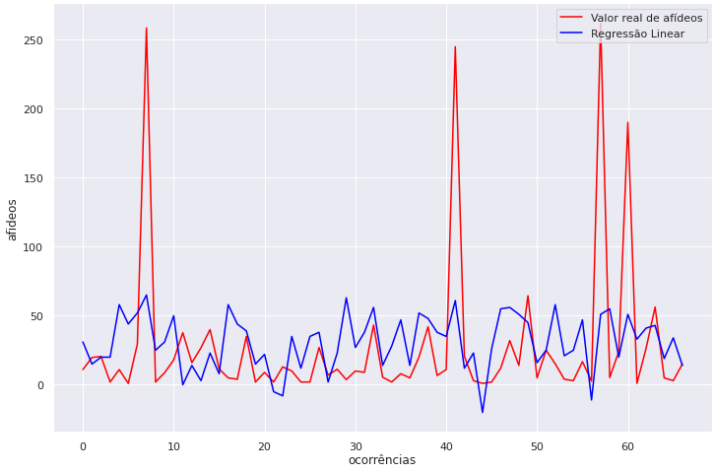


Figura 8 – Ocorrências de afídeos em comparação à predição do modelo.

3.5.3 Redes Neurais Artificiais

Para implementar a técnica de Redes Neurais Artificiais foi utilizada a classe *MLPRegressor* da *Scikit-learn*. Para o treinamento do modelo foram feitas algumas configurações que tiveram como base os resultados do recurso *GridSearchCV* e testes realizados de forma manual.

Ao fim do processo de testes para definir a melhor configuração, ficou definido que a rede deve possuir quatro camadas de 8 neurônios cada e o número máximo de iterações no processo de aprendizagem é de no máximo 5000 passos.

Como forma de visualizar as predições desta técnica, a figura 9 mostra ocorrências dos dados de teste, cada ocorrência possui o valor real de afídeos (linha vermelha) e o valor predito pelo modelo (linha azul). O eixo X representa as ocorrências e o eixo Y a quantidade de afídeos. A acurácia obtida com coeficiente de determinação R^2 é de 26,4%.

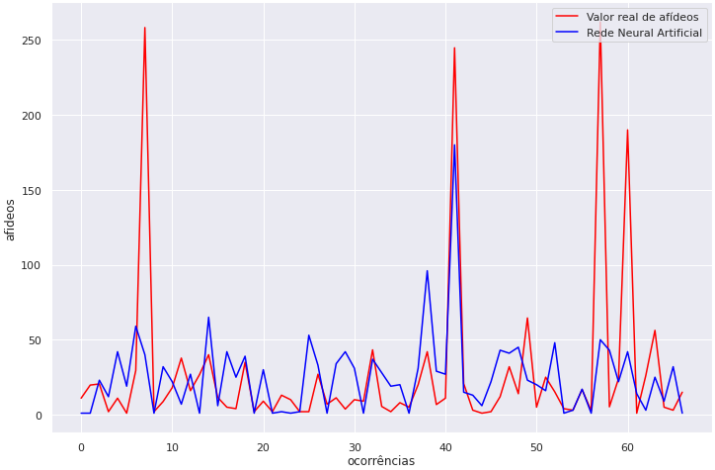


Figura 9 – Ocorrências de afídeos em comparação à predição do modelo.

3.5.4 Árvore de decisão

A árvore de decisão foi implementada utilizando a classe *DecisionTreeRegressor*. Com base no que disse Lan et al. (2020, p. 2053), "o crescimento irrestrito da árvore de decisão pode facilmente levar ao sobreajuste e degradação de sua capacidade de generalização".

Sendo assim, para o processo de aprendizagem da árvore utilizou-se uma técnica conhecida como poda, que limita a profundidade máxima que a árvore terá, evitando assim a ocorrência do chamado *overfitting*. Após a realização de diversos testes manuais, definiu-se como oito a profundidade máxima para árvore de decisão.

A figura 10 mostra ocorrências da base de dados de testes, cada ocorrência possui o valor real de afídeos (linha vermelha) e o valor predito pelo modelo (linha azul). O eixo X representa as ocorrências e o eixo Y a quantidade de afídeos. Sobre a base de dados para teste este modelo chegou a acurácia de 29,3%.

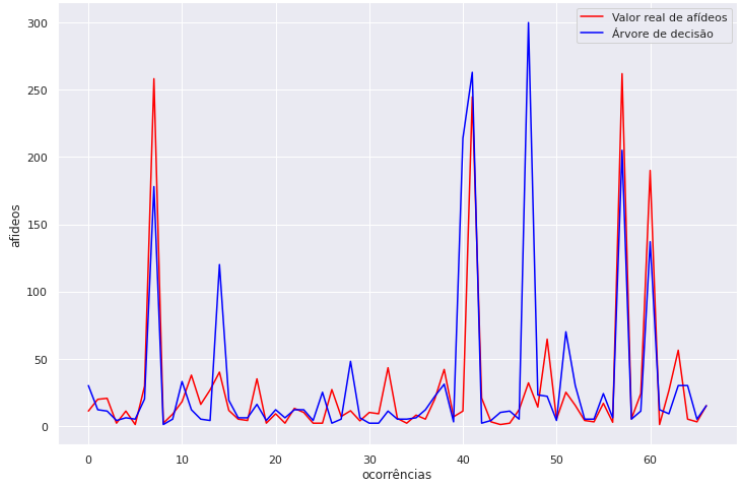


Figura 10 – Ocorrências de afídeos em comparação à predição do modelo.

3.5.5 Floresta aleatória

A classe resposável por implementar a floresta aleatória na biblioteca *Scikit-learn* é a *RandomForestRegressor*. Por implementar o conceito de árvore, com auxilio do recurso *GridSearchCV* chegou-se a conclusão de que para o treinamento do modelo seria limitada a sete a profundidade máxima de cada uma das árvores.

Também definiu-se como dez o número máximo de árvores presentes na floresta, seis para o mínimo de amostras necessárias para dividir um nó interno e dois para o número mínimo de amostras para estar em um nó folha.

Na figura 11 é possível visualizar as ocorrências da base de dados de testes, cada ocorrência possui o valor real de afídeos (linha vermelha) e o valor predito pelo modelo (linha azul). O eixo X representa as ocorrências e o eixo Y a quantidade de afídeos. A acurácia da floresta aleatória sobre os dados separados para testes alcançou o valor de 41,4%.

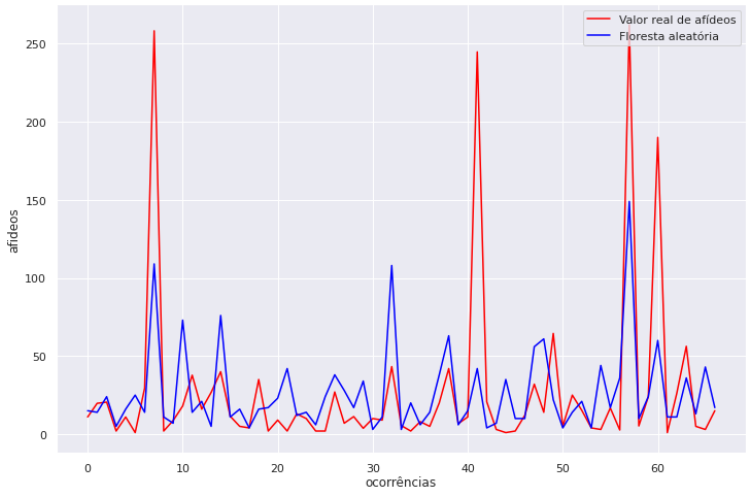


Figura 11 – Ocorrências de afídeos em comparação à predição do modelo.

3.6 Análise dos modelos

Para entender de que forma os dados foram utilizados pelos modelos para realizar as predições, a presente equação refere-se ao modelo de regressão linear desenvolvido. Nela, o coeficiente linear da reta é representado pela constate -65,79 enquanto os valores do coeficiente angular correspondem respectivamente aos campos de: precipitacao (-0,18), semana (0,26), temperaturaMaxima (5,89), temperaturaMinima (-6,88), frqChuva (8,15) e fenomeno (18,50).

$$\gamma = -65,79 - 0,18 \times_1 +0,26 \times_2 +5,89 \times_3 -6,88 \times_4 +8,15 \times_5 +18,50 \times_6 \qquad (3)$$

Através da análise da equação, acredita-se que o valor coeficiente linear da reta explica a predição de valores negativos para a população de afídeos, que podem ser observados na figura 8. Também é possível analisar a intensidade e sentido que o modelo previu para cada uma das variáveis utilizadas.

Para as técnicas de árvore de decisão e floresta aleatória existe um recuros que retorna a importância das variáveis utilizadas durante o treinamento, sendo que este é exibido em formato de porcentagem (SCIKIT-LEARN. . . , 2021). Porém, para a técnica de rede neural artificial, nenhum recurso similar foi encontrado na documentação da biblioteca *Scikit-learn*.

Sendo assim, a árvore de decisão classificou como importância das variáveis: 22% para precipitacao; 17% para semana; 06% para temperaturaMaxima; 34% para temperaturaMinima; 09% para frqChuva; 12% para fenomeno. Enquanto a técnica de floresta aleatória definiu como porcentagem de importância durante o treinamento: 10% para precipitacao; 35% para semana; 14% para temperaturaMaxima; 20% para temperaturaMinima; 08% para frqChuva; 13% para fenomeno.

Como forma de comparar os resultados dos modelos desenvolvidos, a figura 12 mostra uma tabela contendo as dez primeiras ocorrências dos dados separados para teste (Valor real de afídeos) e os valores obtidos pelos modelos de predição desenvolvidos.

	Valor real de afídeos	Árvore de decisão	Rede Neural Artificial	Floresta Aleatória	Regressão Linear
0	11.00	30.0	1.0	15.0	31.0
1	19.75	12.0	1.0	14.0	15.0
2	20.50	11.0	23.0	24.0	20.0
3	2.00	4.0	12.0	5.0	20.0
4	11.00	6.0	42.0	16.0	58.0
5	1.00	5.0	19.0	25.0	44.0
6	29.50	20.0	59.0	14.0	52.0
7	258.25	178.0	40.0	109.0	65.0
8	2.00	1.0	1.0	11.0	25.0
9	8.75	5.0	32.0	7.0	31.0

Figura 12 – Populações de afídeos em comparação a inferências realizadas pelos modelos.

Se tem como destaque as técnicas de Árvore de decisão e Floresta aleatória que, salvo excessões, foram capazes de acompanhar a dinâmica da população de afídeos, inclusive, na ocorrência da população de afídeos acima da média que aconteceu no registro de índice 7. Por outro lado, as técnicas de Redes neurais artificiais e Regressão linear obtiveram resultados inferiores quando comparadas às outras duas técnicas estudadas.

A figura 13 apresenta a base de dados de testes completa em comparação aos valores preditos pelos modelos. Nela é possível análisar o mesmo padrão para Floresta aleatória e Árvore de decisão, sendo que este último se mantem mais próximo da real ocorrência de afídeos quando comparada à Floresta aleatória. Outro fator observado é a dificuldade apresentada pelas técnicas de Regressão linear e Redes Neurais Artificiais em prever a ocorrência da população de afídeos acima da média.

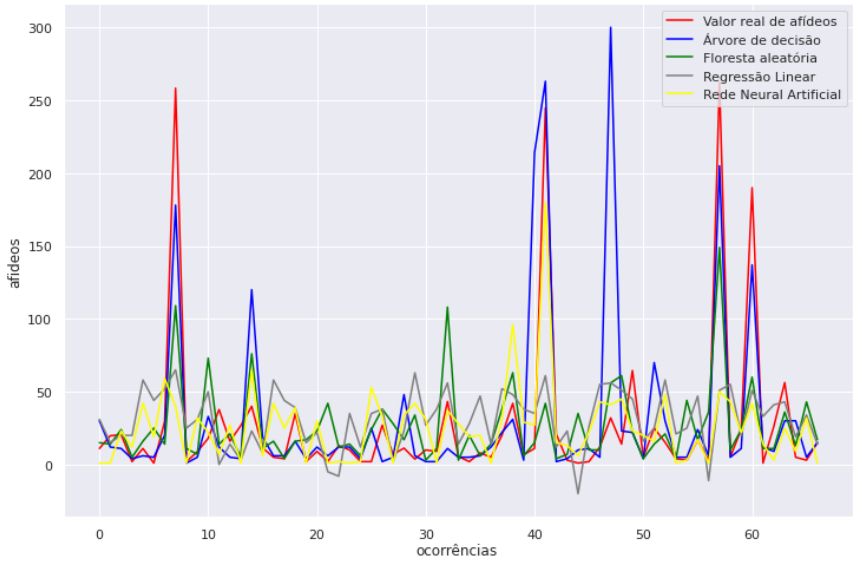


Figura 13 – Ocorrências de afídeos em comparação à predição do modelo.

A técnica que apresentou maior acurácia no coeficiente de determinação R^2 foi Floresta aleatória (41,4%). Porém tanto Floresta aleatória quanto Árvore de decisão apresentaram bons resultados de predições.

Para analisar individualmente o comportamento dos modelos durante o ano todo, na figura 14 são plotados gráficos para cada cidade e ano estudado. Nestes gráficos a

linha vermelha representa a população de afídeos que de fato aconteceu e as demais linhas os valores preditos pelos modelos. O eixo X representa as semanas do ano e o eixo Y a quantidade de afídeos. A cidade e ano análisado é exposto acima de cada um dos gráficos.

Além da aproximação da população real com a predita, foi levado em consideração a capacidade que as técnicas estudadas tem de se adaptar aos períodos de baixa na população de afídeos, junho e agosto (semanas 23 à 32 de cada ano) e os períodos de população de afídeos acima da média, entre os meses de setembro à outubro (semanas do ano 36 à 42).

Figura 14 – Ocorrências semanais de afídeos em comparação à predição do modelo.

Conforme calculado pelo coeficiente de determinação R^2, o modelo de regressão linear é o que apresenta o pior desempenho entre as técnicas abordadas. Em períodos normais para população de afídeo seu desempenho é razoável, porém o modelo não é capaz de prever diminuição ou aumento na população dos insetos. Também em diversos momentos o modelo prevê populações negativas de afídeos, o que inviabiliza sua aplicação.

Com exceções, a rede neural artificial foi capaz de prever alguns picos na população de afídeos, porém se mostrou inconstante em períodos onde a dinâmica da população destes insetos se mantém dentro da média. Apesar de acontecer com menos frequência quando comparado ao modelo de regressão linear, a rede neural artificial também prevê alguns casos onde a população de afídeo é negativa.

Os modelos de árvore de decisão e floresta aleatória são os que melhor representam os valores reais da população de afídeos. Ambos são capazes de identificar a redução populacional que estes insetos sofrem durante os meses de junho à agosto, bem como seu aumento entre os meses de setembro à outubro.

O modelo de Árvore de decisão aparenta ter maior aproximação com as ocorrências reais de afídeos quando comparado ao modelo de floresta aleatória, porém ambos são eficazes para predizer os dados estudados.

Acredita-se ainda ser possível melhorar os resultados dos modelos com a adição de novas variáveis climáticas para o treinamento, pois, conforme citado por PEREIRA, BANZATTO e FURIATTI (1999), variáveis como umidade relativa do ar, insolação, vento, balanço hídrico, radiação solar global também possuem correlação com a dinâmica populacional destes pulgões, porém não estão sendo abordadas neste estudo. Neste mesmo sentido variáveis como parasitoides e predadores, que são inimigos naturais dos afídeos (MELLO, 1994) também devem melhorar a acurácia dos modelos treinados por também apresentarem correlação com a dinâmica populacional dos afídeos.

4 Considerações finais

O objetivo de organizar uma base de dados referente à regiões onde existem monitoramento de afídeos foi alcançado. O que possibilitou o estudo da correlação existente entre população de afídeos e dados climáticos, a fim de extrair conhecimentos sobre a dinâmica populacional dos afídeos com base em dados e fenômenos climáticos.

Através da matriz de correlação e também a visualização dos gráficos foi possível verificar a forma de como a alteração nas variáveis climáticas de temperatura, precipitação e fenômenos climáticos como El niño e La niña impactam na quantidade populacional destes pulgões. Conforme visto em alguns estudos citados nos resultados, também foi possível verificar períodos do ano em que as populações de afídeos aumentam (setembro e outubro) e diminuem (junho, julho e agosto).

Ainda através da análise de dados realizada no trabalho, este estudo propõe que a quantidade de dias de chuva que ocorrem em uma determinada semana também é um fator que possui correlação com a diminuição ou aumento na população de afídeos.

Quanto aos modelos desenvolvidos conclui-se que as técnicas que utilizam Árvore de decisão foram capazes de predizer os dados estudados com maior eficiência quando comparados às técnicas de Regressão Linear e Redes Neurais Artificiais. Este fator pode ser verificado também através do coeficiente de determinação R^2 aplicado sobre os dados separados para teste, onde Regressão Linear apresentou uma acurácia de 11,4%; Redes Neurais Artificiais a acurácia de 26,4%; Árvore de decisão acurácia de 29,3% e Floresta aleatória acurácia de 41,4%.

Como trabalhos futuros, acredita-se que a acurácia dos modelos pode ser melhorada com a utilização de um conjunto maior de dados para o processo de treinamento, podendo estes dados serem referentes a mais localidades em que se tem o monitoramento de afídeos.

Outra forma de melhorar os resultados de predição deve ser através da adição de novas variáveis que influênciam na população de afídeos como por exemplo: variáveis climáticas como umidade relativa do ar, insolação, vento, entre outras; inimigos naturais dos afídeos como parasitoides e predadores.

Outro fator que deve ser levado em consideração para trabalhos futuros é a realização da discretização dos valores, que significa transformar os dados contínuos que foram estudados para dados categóricos, que abranjam determinados períodos de dados, podendo assim melhorar a assertividade dos modelos e se manter útil para a predição dos valores populacionais dos afídeos.

Por fim, com base nas considerações citadas, pode-se concluir que o objetivo de extrair conhecimentos sobre a dinâmica populacional dos afídeos em relação às variáveis climáticas foi alcançado, pois com o trabalho foi possível detectar situações em que a população de afídeos aumenta ou diminui conforme a variação dos dados climáticos estudados.

Para as técnicas de predição utilizadas, levando-se em consideração a análise realizada nos gráficos e os resultados obtidos no coeficiente de determinação R^2, técnicas que implementam Árvores de decisão como a própria Árvore de decisão e a Floresta aleatória são mais apropriados para o tipo de dados estudado neste trabalho.

Referências

AALST, W. V. D. Data science in action. In: *Process mining*. [S.l.]: Springer, 2016. p. 3–23. Citado na página 2.

ABRAHAM, A. et al. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, Frontiers, v. 8, p. 14, 2014. Citado na página 6.

AUAD, A. M.; FREITA, S. d.; BARBOSA, L. R. Ocorrência de afídeos em alface (lactuca sativa l.) em cultivo hidropônico. *Neotropical Entomology*, SciELO Brasil, v. 31, n. 2, p. 335–339, 2002. Citado na página 3.

BORGES, L. E. *Python para desenvolvedores: aborda Python 3.3*. [S.l.]: Novatec Editora, 2014. Citado na página 6.

CUNHA, G. d. El niño oscilação sul e perspectivas climáticas aplicadas no manejo de culturas no sul do brasil. *Revista Brasileira de Agrometeorologia*, v. 7, n. 2, p. 277–284, 1999. Citado na página 4.

CUNHA, G. R. da et al. El niño-oscilaçao do sul e seus impactos sobre a cultura de cevada no brasil. 2001. Citado 3 vezes nas páginas 2, 4 e 10.

DATE, C. J. *Introdução a sistemas de bancos de dados*. [S.l.]: Elsevier Brasil, 2004. Citado na página 4.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. *Acta Scientiarum. Technology*, v. 24, p. 1715–1725, 2002. Citado na página 11.

FERNANDES, G. A. G.; ARAúJO, H. L.; GOMES, R. C. C. Sistema automatizado de aquisição de dados meteorológicos. campus Santa, Jornada Acadêmica da UEG, 2013. Citado na página 3.

FINKLER, C. L. L. Controle de insetos: uma breve revisão. *Anais da Academia Pernambucana de Ciência Agronômica*, v. 8, p. 169–189, 2011. Citado na página 1.

GALDINO, T. V. et al. Efeito de elementos climáticos e inimigos naturais sobre a densidade de pulgões na cultura do repolho. 2007. Citado na página 14.

GALLO, D. et al. *Entomologia Agrícola*. Piracicaba, São Paulo, Brasil: FEALQ, 2002. Citado na página 3.

GASSEN, D. N. Insetos associados a cultura do trigo no brasil. *Embrapa Trigo-Circular Técnica (INFOTECA-E)*, Passo Fundo: EMBRAPA-CNPT, 1984., 1984. Citado na página 2.

INPE - Instituto Nacional de Pesquisas Espaciais. 2021. <http://www.inpe.br/>. Accessed: 2021-06-13. Citado na página 3.

KAMILARIS, A.; KARTAKOULLIS, A.; PRENAFETA-BOLDÚ, F. X. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, Elsevier, v. 143, p. 23–37, 2017. Citado na página 2.

LAN, T. et al. A comparative study of decision tree, random forest, and convolutional neural network for spread-f identification. *Advances in Space Research*, Elsevier, v. 65, n. 8, p. 2052–2061, 2020. Citado 3 vezes nas páginas 5, 6 e 16.

LAZZARETTI, A. T. et al. Agrodb–integration of database management systems with crop models. 2016. Citado na página 4.

LAZZARETTI, A. T. et al. Trapsystem-uma aplicação para gerenciamento de dados coletados a partir de armadilhas de insetos. In: IN: REUNIÃO DA COMISSÃO BRASILEIRA DE PESQUISA DE TRIGO E TRITICALE, 10 .... *Embrapa Trigo-Artigo em anais de congresso (ALICE)*. [S.l.], 2016. Citado na página 7.

LIDIANE, L. M. K. et al. Análise fatorial por meio da matriz de correlação de pearson e policórica no campo das cisternas. *E&S Engineering And Science*, v. 7, n. 1, p. 58–70, 2018. Citado na página 4.

MAFIA, R. G. et al. Variáveis climáticas associadas à incidência de mofo-cinzento em eucalipto. *Fitopatologia Brasileira*, SciELO Brasil, v. 31, n. 2, p. 152–157, 2006. Citado na página 3.

MCKINNEY, W. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, Seattle, v. 14, n. 9, p. 1–9, 2011. Citado na página 7.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. [S.l.]: Novatec Editora, 2019. Citado na página 7.

MELLO, M. E. F. d. Afídeos (homoptera: Aphididae) e seus inimigos naturais em olerícolas, piraquara, paraná. 1994. Citado na página 20.

NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, v. 16, n. 1, 2018. Citado na página 11.

OLIVEIRA, A. M. de. Observações sobre a influência de fatores climáticos nas populações de afídeos em batata. *Pesquisa Agropecuária Brasileira*, v. 6, n. 1, p. 163–172, 1971. Citado 2 vezes nas páginas 2 e 12.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 6.

PEREIRA, A. B.; BANZATTO, D. A.; FURIATTI, R. Influência de elementos climáticos na flutuação populacional de alados de myzus persicae (sulzer)(homoptera: Aphididae). *Revista Brasileira de Agrometeorologia*, v. 7, n. 2, p. 219–225, 1999. Citado na página 20.

PEREIRA, P. d. S.; LAU, D.; JÚNIOR, A. M. Dinâmica populacional de afídeos vetores de bydy: impactos ao rendimento de grãos em trigo. In: N: REUNIÃO DA COMISSÃO BRASILEIRA DE PESQUISA DE TRIGO E TRITICALE, 10 .... *Embrapa Trigo-Artigo em anais de congresso (ALICE)*. [S.l.], 2016. Citado 2 vezes nas páginas 12 e 13.

RODRIGUES, R. L.; MEDEIROS, F. P. D.; GOMES, A. S. Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2013. v. 24, n. 1, p. 607. Citado 3 vezes nas páginas 4, 5 e 11.

ROSENZWEIG, C. et al. Climate change and extreme weather events-implications for food production, plant diseases, and pests. 2001. Citado 2 vezes nas páginas 4 e 13.

SCIKIT-LEARN Machine Learning in Python. 2021. <https://scikit-learn.org/stable/>. Accessed: 2021-06-13. Citado 2 vezes nas páginas 6 e 17.

SOLAGNA, E. A. Análise de perdas de pacotes de uma rede lorawan. Universidade de Passo Fundo, 2020. Citado na página 4.

SOUSA, Á. Coeficiente de correlação de pearson e coeficiente de correlação de spearman: o que medem e em que situações devem ser utilizados? *Correio dos Açores*, Gráfica Açoreana, Lda, p. 19–19, 2019. Citado 3 vezes nas páginas 4, 13 e 14.

STERN, D. L. Aphids. *Current Biology*, Elsevier, v. 18, n. 12, p. R504–R505, 2008. Citado na página 3.

TOEBE, J. Um modelo baseado em agentes para o ciclo de vida dos insetos: aplicação na interação afídeo-planta-vírus. Universidade de Passo Fundo, 2014. Citado na página 2.

TOMÉ, A. et al. Dinâmica das populações de afídeos de cereais de inverno em coxilha/rs entre 2011 e setembro de 2013. In: IN: MOSTRA DE INICIAÇÃO CIENTÍFICA DA EMBRAPA TRIGO, 8.; MOSTRA DE PÓS .... *Embrapa Trigo-Resumo em anais de congresso (ALICE)*. [S.l.], 2013. Citado 2 vezes nas páginas 12 e 13.

TORRES, J. D. et al. Aquisição de dados meteorológicos através da plataforma arduino: construção de baixo custo e análise de dados. *Scientia Plena*, v. 11, n. 2, 2015. Citado na página 3.

VENDRUSCOLO, D. G. S. et al. Estimativa da altura de eucalipto por meio de regressão não linear e redes neurais artificiais. *Revista Brasileira de Biometria*, v. 33, n. 4, p. 556–569, 2015. Citado na página 5.

WIEST, R. Crescimento populacional de rhopalosiphum padi (linnaeus, 1758)(hemiptera: Aphididae) em trigo: uma aproximação por simulação. Universidade de Passo Fundo, 2019. Citado 3 vezes nas páginas 2, 3 e 7.