UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA "LUIZ DE QUEIROZ"

Desempenho de algoritmos de aprendizado de máquina na predição de áreas com cana-de-açúcar

Aluna: Ana Clara Arantes Villas Bôas de Barros
Orientador: Prof. Dr. Marcelo Andrade da Silva

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Engenheiro Agrônomo.

Piracicaba
2023

Ana Clara Arantes Villas Bôas de Barros

Desempenho de algoritmos de aprendizado de máquina na predição de áreas com cana-de-açúcar

Orientador:
Prof. Dr. Marcelo Andrade da Silva

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Engenheiro Agrônomo.

Piracicaba
2023

Sumário

Resumo

O presente trabalho tem como objetivo a comparação de modelos de aprendizado de máquinas para a identificação da cana-de-açúcar em imagens Landsat no estado de São Paulo, um dos maiores estados produtores de cana no Brasil (CONAB, 2020). As covariáveis foram utilizadas à três modelos de aprendizado de máquinas: Regressão Logística (com e sem penalização), Árvores de Decisão e Florestas aleatórias. Preocupados com a escalabilidade do uso do modelo, decidimos aplicar a análise de componentes principais (PCA) para reduzir a dimensionalidade dos dados. Concluímos pela análise PCA que os dois componentes responsáveis pela maior parte da variabilidade do modelo dizem respeito às variáveis que são extremamente correlacionadas ao período das águas ou à estação seca. Neste estudo, a aplicação das novas covariáveis, produzidas pelo PCA, aos modelos significou uma redução de 80% do tempo computacional despendido no processo. Quanto à acurácia dos modelos, o melhor modelo após redução da dimensionalidade foi o Florestas aleatórias, com uma acurácia de 77,00%, seguido pelo modelo de Árvore de Decisão, com 70,04%, e pelo modelo de Regressão Logística, com 68,58%.

Palavras-chaves: Sensoriamento remoto, aprendizado de máquinas, predição da cobertura do solo, análise de componentes principais.

Abstract

The present work aims to compare machine learning models in the identification of sugarcane with Landsat images in the state of São Paulo, one of the largest sugarcane-producing states in Brazil (CONAB, 2020). The covariates were used in three machine learning models models: Logistic Regression (with and without regularization), Decision Trees, and Random forest. Concerned about the scalability of the model's use, we decided to apply Principal Component Analysis (PCA) to reduce the dimensionality of the data. Through PCA, we concluded that the two components responsible for most of the model's variability are related to variables that are highly correlated with the wet season or the dry season. In this study, the application of the new covariates produced by PCA to the models resulted in an 80% reduction in the computational time required for the process. Regarding the accuracy of the models, the best model after dimensionality reduction was the Random forest, with an accuracy of 77.00%, followed by the Decision Tree model with 70.04%, and the Logistic Regression model with 68.58%.

Keywords: Remote sensing, machine learning, principal component analysis (PCA).

1 Introdução

A cana-de-açúcar tem grande importância no agronegócio e consequentemente na economia nacional, já que o Brasil é o maior produtor de açúcar e um dos grandes mercados mundiais de biocombustível (OECD-FAO, 2018; Luciano et al., 2019). Essa cultura tem sua produção disseminada por todo o país, com uma maior concentração na região sudeste, sendo o estado de São Paulo o seu principal produtor (CONAB, 2020). De acordo com Nonato e Oliveira (2013), nos últimos anos, estudos sobre o uso de dados de sensoriamento remoto têm sido amplamente realizados no monitoramento de cana-de-açúcar, gerando grandes avanços e inovações tecnológicas na automação da identificação desse cultivo, assim como no desenvolvimento de metodologias de mapeamento da cobertura do solo. Tradicionalmente, o mapeamento da cana-de-açúcar era realizado por identificação visual, sendo necessário mais tempo de execução do processo e um grande grupo de profissionais altamente treinados (Vieira et al., 2012). Os modelos preditivos podem ser aplicados em diferentes áreas do conhecimento. A seguir, alguns exemplos de aplicações nas Ciências Agrárias são apresentados.

1. **Previsão de produtividade:** usado para estimar a produtividade de uma cultura durante um período determinado. Como mencionado em Oliveira (2010), os modelos de previsão de produtividade são "importante ferramenta de assistência à tomada de decisões para viabilizar sistemas racionais de produção. Apesar da complexidade envolvida na construção desses modelos, os esforços são compensados em função de sua grande aplicabilidade".

2. **Identificação de daninhas:** Os sistemas de identificação de daninhas estão atreladas a implementos agrícolas, que são responsáveis pela aplicação de agroquímicos ou *laser*. A aplicação desses sistemas ajuda a evitar contaminação ambiental e aumenta a segurança dos operadores (Becker et al., 2021).

3. **Análise de mercado:** diferentes modelos podem ser aplicados para estimar a flutuação dos preços de *comodities*. Dos modelos de predição, os mais utilizados nesse ramo são *forecasting*, redes neurais ou uma abordagem mista das duas (Shahwan e Odening, 2007).

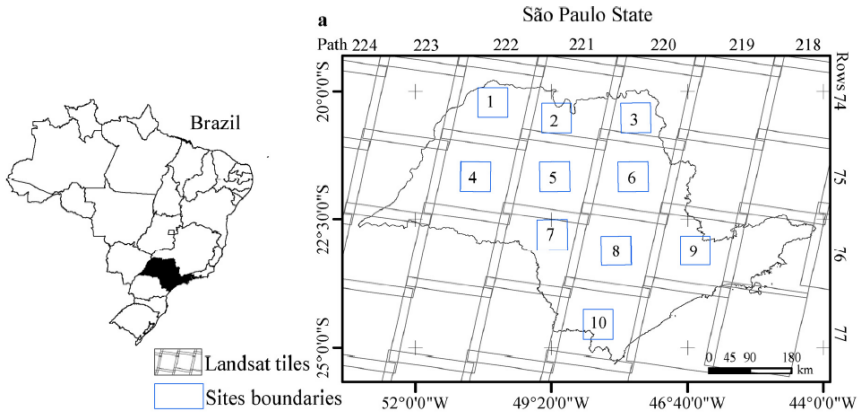
Outra aplicação de modelos preditivos bastante comum em Ciências Agrárias é a classificação do uso do solo, pois esses modelos são essenciais para o mapeamento e a tomada de decisão (da Costa Araújo Filho et al., 2007). É comum que algumas instituições e órgãos públicos mapeiem regiões ou áreas específicas quanto ao uso da terra, mas o mapeamento em larga escala é trabalhoso e resulta em um alto custo de aplicação. Dessa forma, os dados obtidos por satélites se apresentam como uma ferramenta muito atrativa para a identificação dessas áreas (Dainese, 2001). Em contraste ao mapeamento de aspectos físicos do ambiente, o mapeamento de culturas está suscetível a constantes mudanças. Logo, para um monitoramento dessas áreas é necessário que haja o mapeamento frequente. Assim, a predição do uso da terra através de imagens de satélite possibilita uma identificação rápida e que pode ser repetida periodicamente, sendo ideal para o monitoramento em grande escala.

Para alimentar os modelos de predição para a classificação do uso da terra foram utilizados neste estudo bandas e índices do satélite Landsat, compilados durante um período de 17 meses (Luciano et al., 2018). Esses dados foram submetidos ao preenchimento de valores faltantes por meio de interpolação e à correção atmosférica antes de serem incorporados ao banco de dados.

A área de estudo selecionada para este trabalho compreende os municípios de São Carlos, Araraquara, Santa Lúcia, Rincão, Motuca, Luís Antônio, Descalvado, Pirassununga e Ibaté, situada na área 6, conforme a Figura 1.

Os dados coletados na região selecionada e aplicados neste estudo ilustram a importância do sensoriamento remoto e da automatização da classificação de áreas cultivadas, possibilitando o ganho de informações atualizadas de forma ágil e, consequentemente, permitindo a tomada de decisão e a formulação de políticas regionais.

O objetivo deste estudo é comparar a qualidade da predição de três importantes modelos de aprendizado de máquina na identificação de áreas que possuem plantações de



Fonte: Luciano et al. (2019)

Figura 1: Áreas de pesquisa. A região escolhida para esse estudo é a região número 6, que contempla cidades como Araraquara e São Carlos.

cana-de-açúcar utilizando imagens Landsat: Regressão Logística, Árvore de Classificação e Florestas aleatórias. Além disso, a análise de componentes principais é utilizada para reduzir o número de covariáveis utilizadas nos modelos. Para isso, foram selecionados índices de vegetação e bandas espectrais obtidos a partir das imagens do satélite Landsat em uma região com grande concentração de área plantada com cana-de-açúcar no estado de São Paulo.

2 Revisão de literatura

2.1 Cana-de-açúcar

A cana-de-açúcar é de grande importância no agronegócio e na economia brasileira, visto que o Brasil é o maior produtor mundial dessa cultura. Ao considerar para os produtos derivados na cana, o Brasil também se destaca. O país é um dos maiores mercados de açúcar e biocombustível no cenário mundial. Dentro do Brasil, a região sudeste é a maior produtora, na região os maiores destaques são para os estados de São Paulo e Minas Gerais (CONAB, 2020).

A cana de açúcar é uma cultura proeminente no Brasil há muito tempo. Logo no final do século XVIII, a cana-de-açúcar passou a ganhar destaque na economia brasileira, sendo uma das principais formas de geração de riqueza na colônia portuguesa juntamente com a extração de minérios. Com o passar do tempo, o setor sucroenergético foi sendo mecanizado, dando origem aos engenhos. Estes engenhos e a progressiva diminuição do trabalho exploratório levaram a produção da cana e de seus subprodutos a ser mais efetiva e, consequentemente, mais lucrativa. (ARAÚJO e Santos, 2013) Após a independência do Brasil, e 1822, o país já tinha posição de destaque no abastecimento mundial do mercado de açúcar e, com a crise do petróleo que se iniciou em 1973, o Brasil aproveitou a oportunidade para desenvolver um biocombustível derivado da cana, o etanol (Nocelli et al., 2017).

Biocombustíveis são combustíveis gerados a partir de fontes renováveis, como a cana, as oleaginosas e até os microrganismos. O uso de combustíveis fósseis implicam em dois problemas principais, a eventual escassez desses produtos e também a produção de gases de efeito estufa (Peres et al., 2005). Os combustíveis de fontes renováveis além de reduzirem emissão de gases de efeito estufa reduzem a dependência dos combustíveis fósseis, além disso ainda criam emprego no meio rural (Silva e Konradt-Moraes, 2012). Atualmente, a cana-de-açúcar é considerada não apenas uma importante cultura na economia brasileira, mas também uma fonte de biomassa energética. O uso desse tipo de

tecnologia tem alta relevância no cenário nacional e internacional para diversas áreas, além de aumentar a oferta de emprego também ajuda no desenvolvimento econômico do país (Rodrigues, 2010).

2.2 Sensoriamento remoto

O sensoriamento remoto é um processo em que se obtém informações sobre o alvo sem que haja contato direto com ele. Esse processo também é responsável pela interpretação dos dados obtidos, sejam eles imagens ou outra forma de dados, como informações de sensores e índices. Para que esse processo possa ocorrer, primeiramente, precisamos de uma fonte de energia. Essa energia, no caso de imagens de satélite, é o sol ou a energia em forma de calor vindo da terra. Em seguida, essa energia é emitida ou retransmitida para a atmosfera, e então precisamos de um sensor que possa captar essa energia (Lillesand et al., 2015). A seguir, temos alguns tipos de sensores .

1. **Sensores ópticos:** utilizam-se de luz ou ondas eletromagnéticas para obter informações sobre o alvo. As informações obtidas podem ser combinadas para facilitar a interpretação das imagens, como a combinação das bandas vermelha, verde e azul que gera o conhecido RGB (Red-Green-Blue) que cria variações de cores visíveis aos humanos.

2. **Sensores térmicos:** esse tipo de sensor, como o nome sugere, é capaz de aferir a temperatura do alvo através de radiação infravermelha.

Depois que a energia é captada, é possível obter as informações captadas pelo sensor digitalmente. Assim, essas informações são interpretadas e podem ser disponibilizadas em forma de mapas ou planilhas.

O sensoriamento remoto surgiu por volta de 1858, com a primeira foto aérea tirada por balão. Desde então, a tecnologia avançou rapidamente. Atualmente, o sensoriamento remoto pode ser aplicado em diversas áreas do conhecimento. Na área das Ciências Agrárias, passou a ser utilizado a partir da década de 1930, mas foi somente

em 1937 que as fotografias aéreas de algumas partes dos Estados Unidos começaram a ser tiradas recorrentemente (Lillesand et al., 2015). Na década de 1970, a NASA (Administração Nacional da Aeronáutica e Espaço, em inglês *National Aeronautics and Space Administration*) lançou o satélite Landsat-1, o primeiro de muitos que seriam lançados. Até hoje temos satélites Landsat em órbita sendo o Landsat-9 o mais recente.

A evolução das geotecnologias com o uso de sensores situados na órbita do planeta permite o desenvolvimento de metodologias para a classificação do uso do solo, em especial de áreas agrícolas. Essas metodologias podem fornecer informações necessárias para diversas aplicações como previsão e monitoramento de safras de forma remota (Nonato e Oliveira, 2013).

Os processos de classificação do uso da terra eram realizados de forma manual por identificação visual das áreas, gerando um grande custo de tempo e mão de obra. Porém, é possível identificar pela análise de dados obtidos a partir de imagens de satélites um comportamento específico da cana-de-açúcar durante seu ciclo, um total de um ano e meio. A metodologia propõe que a área selecionada seja segmentada pelo método Baatz (2000), sendo coletado então os valores dos pixels durante o ciclo. Dessa forma, cada segmento terá uma série temporal com os valores dos seus pixels do início ao fim de seu ciclo, o conjunto das séries temporais de todos os segmentos formará nosso banco de dados. Formado o banco de dados, podemos aplicar ao banco algoritmos de aprendizado de máquinas para predizer quais áreas são plantios de cana-de-açúcar (Vieira et al., 2012).

2.3 Aprendizado de máquina

O aprendizado de máquina surgiu do reconhecimento de padrões e da ideia de que computadores podem aprender a partir de conjuntos de dados para realizar tarefas específicas, identificando padrões e tomando decisões com o mínimo de intervenção humana. Em geral, as máquinas aprendem com computação e dados anteriores para tomar decisões e produzir resultados confiáveis (Morettin e Singer, 2022; Izbicki e dos Santos,

2020).

O termo "computação cognitiva" surgiu na década de 1960 criada por Arthur Samuel enquanto trabalhava na IBM. Anos depois, em 2010, a IBM atualizou o termo para "aprendizado de máquina" como forma de marketing para atrair clientes e novos empregados. Apesar do marketing, não há nada nesse processo de "cognitivo" ou "aprendizado", os modelos de aprendizado de máquina são modelos estatísticos que recebem dados e tiram as características principais para identificar os alvos. Assim, podemos resumir o aprendizado de máquina como uma adaptação dos dados a uma fórmula matemática (Burkov, 2019). Esses modelos de aprendizado de máquina podem ser divididos em:

1. **Supervisionado:** neste caso, usamos um banco de dados para treinar o modelo em que fornecemos a classificação real das observações fornecidas.

2. **Não-Supervisionado:** diferente dos modelos supervisionados, este modelo não recebe a classificação real dos dados. Esse tipo de modelo pode ter diferentes objetivos como a "clusterização" dos dados, onde separamos os dados em grupos com características similares. Outro exemplo é a redução de dimensionalidade, que diminui a quantidade de covariáveis, mantendo suas características mais importantes.

3. **Semi-Supervisionado:** os modelos deste tipo recebem tanto dados com sua classificação real quanto sem classificação, geralmente este segundo está em maior quantidade.

4. **Aprendizado por reforço:** o modelo trabalha em um tipo de ambiente de aprendizado onde é "recompensado" com um *feedback* positivo ou negativo de acordo com a sua resposta.

2.3.1 Regressão logística

Os métodos de regressão são capazes de predizer o resultado baseando-se em covariáveis. Esses modelos usam covariáveis e procuram uma equação que melhor se

adapte ao comportamento dos dados. Uma vez formada, essa equação pode ser aplicada a novos dados para prever seu resultado. Existem diferentes tipos de regressão, cada modelo tem suas condições ideais de aplicação. Por exemplo, o modelo de regressão linear tem um melhor desempenho quando há uma relação próxima à linearidade entre sua variável explicativa e o resultado. Dito isso, é importante olharmos para nossas covariáveis quando falamos de regressão (Bingham e Fry, 2010).

Os modelos de regressão podem ser divididos em dois tipos: simples e multivariados. Os modelos de regressão simples consideram apenas uma covariável para explicar a variável resposta, enquanto os multivariados podem receber uma quantidade n de covariáveis (Bingham e Fry, 2010). A Regressão Logística pode ser aplicada em diversas maneiras no ramo de Ciências Agrárias, incluindo: previsão de doenças em plantas, análise de fatores de produção e na tomada de decisões.

A previsão de doenças em plantas é usada na identificação de condições favoráveis para o aparecimento de doenças. No trabalho de Henderson et al. (2007), além de determinar as variáveis ambientais envolvidas no aparecimento de *Phytophthora infestans*, o trabalho também teve como objetivo tentar prever a gravidade da doença. Neste caso, foi usada a Regressão Logística Binomial para prever a presença (1) ou ausência (0) da doença, e a Regressão Logística Ordinal foi utilizada na predição da gravidade em uma escala de 0 a 4. O modelo binário teve uma acurácia de 67,5%, sensibilidade de 75% e especificidade de 62,5%. O modelo também apontou que o tempo que a planta permanece em temperaturas entre 10^{o}C e 27^{o}C e a precipitação foram fatores importantes para o modelo.

Na análise de fatores de produção podemos entender a importância de cada uma das covariáveis do ambiente de produção. No trabalho de Lad et al. (2022) foram analisados fatores como chuva, nutrientes no solo, umidade e pH. Foram aplicados diversos modelos, entre eles a Regressão Logística com penalização Lasso e Ridge, com o objetivo de extrair as principais características e predizer a produtividade. Neste caso a acurácia

obtida pelo modelo de Regressão Logística foi de 67,57%.

Como exemplo na tomada de decisões agrícolas podemos citar o trabalho de Katarya et al. (2020), que testou diversos modelos de aprendizado de máquinas na recomendação de culturas na Índia, de acordo com as características do ambiente de cultivo. Na recomendação de culturas, a Regressão Logística estava entre um dos modelos testados e apresentou uma acurácia de 78,48% quando aplicado no estado de Uttar Pradesh e 69,35% quando aplicado no estado de Karnataka.

2.3.2 Árvore de decisão

A respeito do modelo de árvore de decisão, há diversas aplicações. No ramo das Ciências Agrárias suas aplicações incluem: previsão de doenças em plantas, classificação do uso do solo e recomendação de manejo.

No trabalho de Meira et al. (2008), é possível observar a aplicação da árvore de decisão na análise da epidemia da ferrugem do cafeeiro. O estudo usou observações sobre dados meteorológicos, carga pendente do cafeeiro e espaçamento entre as plantas para explicar a evolução da doença. Essa evolução foi separada em três categorias: TX1 - redução ou estagnação, TX2 - crescimento moderado e TX3 - crescimento acelerado. O modelo apresentou uma acurácia de 73% considerando todas as três classes, mas olhando individualmente as classes, a classe TX1 foi a que o modelo conseguiu prever melhor acertando cerca de 88% das observações preditas.

Assim como neste trabalho, há outros que também se utilizam deste modelo para predição da classificação do uso do solo. É o caso do trabalho de Delgado et al. (2012), que captou dados provenientes de satélites para predizer áreas com cana-de-açúcar da Fazenda Santa Fé, localizada no estado de Minas Gerais, em cinco datas diferentes. O modelo aplicado neste contexto foi capaz de predizer a área da cana nas datas com uma acurácia de 98%.

No trabalho de Souza et al. (2010), vemos a aplicação do modelo de árvores de decisão para compreender quais fatores são mais importantes para a produtividade da cultura de cana-de-açúcar, entender estes fatores pode ajudar os produtores a separar zonas de manejo, com o objetivo de ajustar seu manejo à cada zona, ajudando a evitar uso excessivo de insumos. Nesse trabalho, o principal fator para separação das classes de produtividade (baixa, média e alta) foi a altitude. O modelo ajustado obteve uma acurácia de 94%.

2.3.3 Florestas aleatórias

O modelo Florestas aleatórias tem uma grande versatilidade e, por isso, tem se tornado um modelo muito popular. Nas Ciências Agrárias, seu uso inclui: predição do preço de *commodities* e predição da produtividade.

Na aplicação de predição de valores de *commodities*, temos o trabalho de Rani et al. (2022) que usou o modelo para predizer os valores do mercado e ajudar pequenos agricultores a receberem valores mais justos pelos seus produtos. O modelo aplicado teve uma acurácia de 95%. Para facilitar o seu uso, os autores criaram uma plataforma, não sendo necessário noções de programação para aplicá-lo.

Um exemplo de predição de produtividade com o modelo de Florestas aleatórias está no trabalho de Prasad et al. (2020), onde o modelo foi aplicado para predizer a produtividade de algodão a nível regional. As observações fornecidas ao modelo eram provenientes de imagens de satélite como o índice NDVI e outros dados meteorológicos como precipitação e temperatura. O modelo foi aplicado para predizer a produtividade em três períodos e a acurácia dos modelos foram as seguintes: setembro 69%, dezembro 60% e em fevereiro com 39%. De acordo com o próprio autor, a queda brusca da produtividade provavelmente ocorreu em função da extrapolação dos dados fora das condições de ambiente conhecidas pelo modelo.

2.4 Redução de dimensionalidade

Trabalhar com um conjunto de dados com muitas variáveis pode não ser uma tarefa fácil devido às altas correlações entre essas variáveis e ao tempo computacional que, em muitos casos, pode ser excessivamente longo. Assim, uma possível solução é a redução de dimensionalidade dos dados, que consiste em transformar dados de um espaço de alta dimensão em um espaço de dimensão menor, de modo que os dados transformados retenham informações acerca da variabilidade dos dados originais, eliminando informações redundantes e simplificando processos subsequentes (Van Der Maaten et al., 2009). A redução de dimensionalidade é comum em áreas de estudos que lidam com grandes números de observações e/ou variáveis e amplamente aplicada na era do *big data*, em que grandes volumes de dados são produzidos diariamente.

Diversas pesquisas em agronomia utilizam técnicas de redução de dimensionalidade dos dados. Por exemplo, Gilbertson e van Niekerk (2017) utilizaram diferentes técnicas de redução de dimensionalidade antes de aplicar os dados de imagens de satélite Landsat-8 em métodos de aprendizado de máquina para a diferenciação de culturas. Ruiz Hidalgo et al. (2021) propõem um método para a redução de dimensionalidade de imagens hiper espectrais de vegetação e cultivos, e Zhai et al. (2020) utilizam um algoritmo de redução de dimensionalidade para mapear a distribuição espacial de três grandes culturas, incluindo milho, arroz e soja no nordeste da China através de imagens de satélite.

3 Material e métodos

No problema de dados de sensoriamento remoto que estamos estudando, a variável resposta de interesse é uma variável de classificação binária que indica se cada segmento de imagem corresponde à plantação de cana-de-açúcar ou não. Sendo assim, apresentamos na sequência o conjunto de dados obtido por satélite, os *softwares* que estamos utilizando no trabalho e a metodologia estatística para o problema de classificação (Mitchell, 1997; Hastie et al., 2009; Izbicki e dos Santos, 2020) explorando os métodos de Regressão Logística, Árvores de Decisão e Florestas aleatórias.

Os códigos elaborados para este trabalho estão disponíveis no github pelo link: https://github.com/AnaArantesBarros/TCC
Já os dados não serão disponibilizados, pois se tratam de dados privados e não possuímos autorização para sua disponibilização pública.

3.1 Descrição dos dados

Para a formação do banco de dados, foi realizada a segmentação de imagens obtidas pelo satélite Landsat. Cada segmento tem um código para sua identificação e foi realizado um levantamento para verificar se esse segmento é uma área de produção de cana-de-açúcar. Para cada um desses segmentos, foram coletados valores mensais obtidos das bandas espectrais e índices de vegetação durante 17 meses e as informações descritas foram unidas em um banco de dados para sua utilização. Os índices e bandas utilizados foram: NDVI (Índice de Vegetação por Diferença Normalizada), EVI (Índice de Vegetação Melhorado), NDWI (Índice de Água de Diferença Normalizada), NDMI (Índice de Umidade de Diferença Normalizada), SWIR1 e SWIR2 (infravermelho de ondas curtas). Com os seis índices e bandas captados durante 17 meses totalizamos 102 covariáveis no banco, que foram observadas em 46.000 segmentos diferentes.

Na Figura 2 abaixo temos uma ilustração para ajudar na compreensão de como foi realizada a formatação do banco de dados.

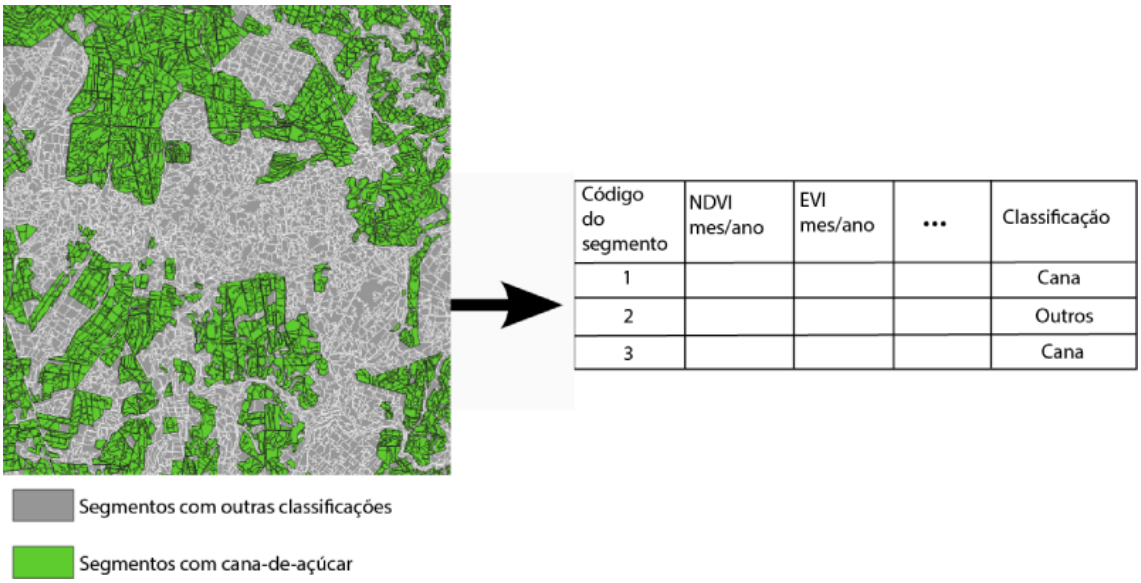


Figura 2: Formação do banco de dados.

3.2 Equipamentos e *softwares* utilizados

O *software* R é uma ferramenta de base aberta, ou seja, toda comunidade pode contribuir para o seu desenvolvimento. O R é a ferramenta mais utilizada pela academia, pois possibilita a computação estatística e gráfica de dados. Por utilizar uma linguagem de programação própria, possibilita uma maior versatilidade ao usuário. Além das funcionalidades básicas do programa, estas ainda podem ser estendidas pelo uso de pacotes desenvolvidos pela comunidade. Dentro do *software* algumas bibliotecas foram necessárias para a aplicação dos modelos, são eles:

1. **ggplot2:** é uma biblioteca responsável pela geração de gráficos. O objetivo de usar essa biblioteca ao invés dos próprios gráficos nativos do R é a sua versatilidade. Além de produzir gráficos esteticamente agradáveis, torna o processo mais interativo, já que possuí diversos padrões prontos. Outra característica importante é a facilidade de combinar múltiplos *layers* em um único gráfico de forma simples (Wickham, 2016).

2. **caret:** o nome "caret" vem da abreviação do inglês *classification and regression training*. Esse pacote é muito útil nas primeiras etapas do aprendizado de máquina, já que ajuda na preparação dos dados de treino e teste para os modelos (Kuhn,

2008).

3. **glmnet:** facilita a aplicação de modelos, como regressão logística sem penalização, regressão logística com penalização lasso ou ridge. Um de seus comandos ajusta o modelo aos dados, enquanto outra usa o modelo ajustado para predizer novas observações (Tay et al., 2023).

4. **rpart:** é um facilitador na aplicação dos modelos de árvores de decisão e alguns métodos de regressão. Além disso, pode gerar um gráfico para as árvores de decisão, permitindo a visualização dos fatores de separação dos dados (Therneau et al., 2022).

5. **FactoMineR:** essa biblioteca permite ao usuário a aplicação de análises multivariadas, como é o caso da análise de componentes principais (PCA). Além disso, também permite a aplicação da análise de correspondência (CA) e da análise de correspondência múltipla (MCA) (Lê et al., 2008).

6. **Factoextra:** permite a visualização dos resultados de análises multivariadas que foram realizadas com a biblioteca "FactoMineR".

Além disso, para facilitar a programação com o software R, usamos um ambiente de desenvolvimento integrado (IDE) chamado Rstudio. O Rstudio é um produto da empresa Posit, e sua versão de base aberta é disponibilizada de forma gratuita.

O QGIS é um Sistema de Informação Geográfica (SIG), de uso profissional e de código aberto, que permite a análise, edição e visualização de dados geográficos (Cutts e Graser, 2018). Essa ferramenta foi criada em 2002 pelo QGIS Development Team e é amplamente utilizado em levantamento de dados georreferenciados.

Neste trabalho, o código no *software* R foi executado em um computador do tipo *desktop* com processor Intel(R) Core(TM) i7-10700, com 16GB de memória RAM e sistema operacional de 64 bits Windows 10.

3.2.1 Regressão logística

O modelo de regressão logística é um caso particular dos modelos lineares generalizados (MLG), em que a variável resposta é binária, isto é, assume apenas os valores 0 ou 1. Esse modelo é muito utilizado em problemas de classificação de diversas áreas do conhecimento com a presença de alta dimensionalidade dos dados Morettin e Bussab (2017).

Sejam Y_i a variável resposta da i-ésima observação e $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})'$ o vetor das d covariáveis da i-ésima observação. Então, o modelo de Regressão Logística pode ser escrito como

$$P(Y_i = 1|\mathbf{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}, \tag{1}$$

em que β_0 é o intercepto do modelo e β_j é o parâmetro associado à j-ésima covariável, com $j = 1, 2, \ldots, d$.

Uma importante propriedade do modelo de Regressão Logística refere-se ao preditor linear da i-ésima observação, $\eta_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}$. Quanto maior o valor de η_i, mais próximo de 1 será a probabilidade dada pela equação 1 e, quanto menor o valor de η_i, mais próximo de 0 será essa probabilidade. Como consequência, um valor positivo para um parâmetro associado a uma determina covariável indica que, quanto maior o valor desta covariável, maior será a probabilidade de ocorrência de sucesso na variável resposta. Por outro lado, um valor negativo indica que essa probabilidade decresce com o aumento dessa covariável. Em suma, os parâmetros do modelo de Regressão Logística possuem fácil interpretação, o que torna o modelo atrativo e muito utilizado.

Para estimar os coeficientes de um modelo de Regressão Logística, podemos utilizar o método de máxima verossimilhança, que consiste em obter os parâmetros que maximizam a função de verossimilhança. É possível também utilizar penalizações para estimar os coeficientes da Regressão Logística, reduzindo a variância do estimador e, consequentemente, obtendo melhor poder preditivo.

3.2.2 Árvore de decisão

Uma Árvore de decisão é construída por particionamentos recursivos no espaço das covariáveis. Cada particionamento recebe o nome de nós e cada resultado final recebe o nome de folha. Inicialmente, o algoritmo verifica se a condição no primeiro nó é satisfeita. Caso seja, segue-se à esquerda. Caso contrário, segue-se à direita. Assim prossegue-se até atingir uma folha.

A criação da estrutura de uma árvore de classificação é feita através de duas grandes etapas: (i) a criação de uma árvore completa e complexa e (ii) a poda dessa árvore, com a finalidade de evitar o super ajuste. Formalmente, a primeira etapa consiste em criar uma partição do espaço das covariáveis em regiões distintas e disjuntas denotadas por $R_1, R_2, \ldots, R_j$. A predição para a resposta Y de uma observação com covariáveis $\mathbf{x}$ que estão em R_k é dada pela moda dos valores da variável resposta das amostras do conjunto de treinamento pertencente àquela mesma região, isto é,

$$g(\mathbf{x}) = \text{moda}\{y_i : \mathbf{x} \in R_k\}.$$

Um critério sugerido para buscar a melhor partição em cada etapa do processo é o índice de Gini, dado por

$$\sum_{R} \sum_{c \in C} \hat{p}_{R,c}(1 - \hat{p}_{R,c}),$$

em que R representa uma das regiões induzidas pela árvore e $\hat{p}_{R,c}$ é a proporção de observações classificadas como sendo da categoria c entre as que estão na região R. Busca-se minimizar esse índice.

Para a etapa da poda, em geral, utiliza-se a proporção de erros no conjunto de validação como estimativa do risco.

3.2.3 Florestas aleatórias

Apesar das Árvores de Decisão serem um método de fácil interpretação e simples entendimento, elas costumam apresentar baixo poder preditivo quando comparadas aos demais estimadores. Para contornar essa limitação, pode-se explorar outro método bastante conhecido chamado de Florestas aleatórias. Essa abordagem consiste obter N

árvores distintas e combinar seus resultados para melhorar o poder preditivo em relação a uma árvore individual. Para criar as N árvores distintas, utiliza-se N amostras *bootstrap* da amostra original. Seja $g_n(\mathbf{x})$ a função de predição obtida segundo a n-ésima árvore. A função de predição dada pelo método é dada por

$$g(\mathbf{x}) = \text{moda}\{g_n(\mathbf{x}), n = 1, 2, \ldots, N\}.$$

3.3 Problemas de classificação

Um problema de classificação é um tipo de problema em aprendizagem de máquina em que o objetivo é atribuir uma categoria para cada observação do conjunto de dados com base em suas características. Os problemas de classificação podem ser binários, quando existem apenas duas classes possíveis, ou não-binários, quando há mais de duas classes. Para a análise de métodos de aprendizado em problemas de classificação, algumas ferramentas, como a matriz de confusão e medidas como acurácia, especificidade e sensibilidade são amplamente utilizadas.

A matriz de confusão é uma tabela que permite visualizar o desempenho de um modelo de classificação. Ela é comumente usada em problemas de aprendizado supervisionado, onde os resultados esperados são conhecidos. A matriz de confusão possui quatro elementos principais: verdadeiro positivo (TP), falso positivo (FP), verdadeiro negativo (TN) e falso negativo (FN).

- Verdadeiro positivo (TP): Representa os casos em que o modelo classificou corretamente uma instância como positiva.
- Falso positivo (FP): Representa os casos em que o modelo classificou incorretamente uma instância como positiva.
- Verdadeiro negativo (TN): Representa os casos em que o modelo classificou corretamente uma instância como negativa.
- Falso negativo (FN): Representa os casos em que o modelo classificou incorretamente uma instância como negativa.

A acurácia é uma métrica comumente utilizada para avaliar o desempenho geral de um modelo de classificação. Ela é definida como a proporção de instâncias corretamente classificadas em relação ao total de instâncias. A fórmula da acurácia é dada por:

Acurácia = (TP + TN) / (TP + FP + TN + FN)

A especificidade (também conhecida como taxa de verdadeiros negativos) é uma medida que indica a proporção de negativos corretamente classificados em relação ao total de negativos. A fórmula da especificidade é dada por:

Especificidade = TN / (TN + FP)

A sensibilidade (também conhecida como taxa de verdadeiros positivos ou recall) é uma medida que indica a proporção de positivos corretamente classificados em relação ao total de positivos. A fórmula da sensibilidade é dada por:

Sensibilidade = TP / (TP + FN)

Essas métricas são importantes para avaliar o desempenho de um modelo de classificação em diferentes aspectos, como a capacidade de identificar corretamente os verdadeiros positivos (sensibilidade) e a capacidade de evitar falsos alarmes (especificidade).

3.4 Análise de componentes principais

A análise de componentes principais (em inglês, Principal Component Analysis - PCA) é uma técnica estatística utilizada para reduzir a dimensionalidade dos dados, preservando a maior parte da informação original. Essa técnica busca identificar os componentes principais, que são combinações lineares das variáveis originais, de forma a maximizar a variância dos dados projetados. Detalhes sobre a PCA podem ser encontrados em Manly e Alberto (2008) e Morettin e Singer (2022).

A PCA é muito utilizada em estudos em que há um grande número de variáveis, cujo objetivo é simplificar a análise, sendo útil para a visualização de dados e o pré-processamento de dados para outras técnicas de análise estatística. De maneira simplificada, a PCA pode ser resumida nos seguintes passos:

1. Padronização dos dados: os dados são padronizados para garantir que todas as variáveis estejam na mesma escala.

2. Cálculo da matriz de covariância ou correlação.

3. Cálculo dos autovetores e autovalores.

4. Determinação do número de componentes principais: essa seleção pode ocorrer com base nos autovalores correspondentes (autovalores maiores indicam que seus componentes principais explicam a maior parte da variabilidade dos dados) e/ou a partir do gráfico "Scree Plot".

5. Projeção dos dados nos componentes principais: os dados originais são projetados nas direções dos componentes principais selecionados no passo anterior. Isso resulta em uma nova representação dos dados em um espaço de menor dimensionalidade.

Ao realizar a PCA, é possível obter uma visão geral dos padrões existentes nos dados, identificar variáveis importantes e reduzir a dimensionalidade dos dados, tornando a análise mais eficiente.

Uma ferramenta bastante útil em PCA é o gráfico "Biplot", que combina informações sobre as observações e as variáveis em um único gráfico, permitindo a análise conjunta desses elementos no espaço de componentes principais. No gráfico "Biplot", cada ponto representa uma observação, e cada vetor (seta) representa uma variável original. Assim, essa ferramenta fornece a similaridade de observações, as correlações entre as variáveis originais, a importância dessas variáveis e informações para interpretar as componentes principais.

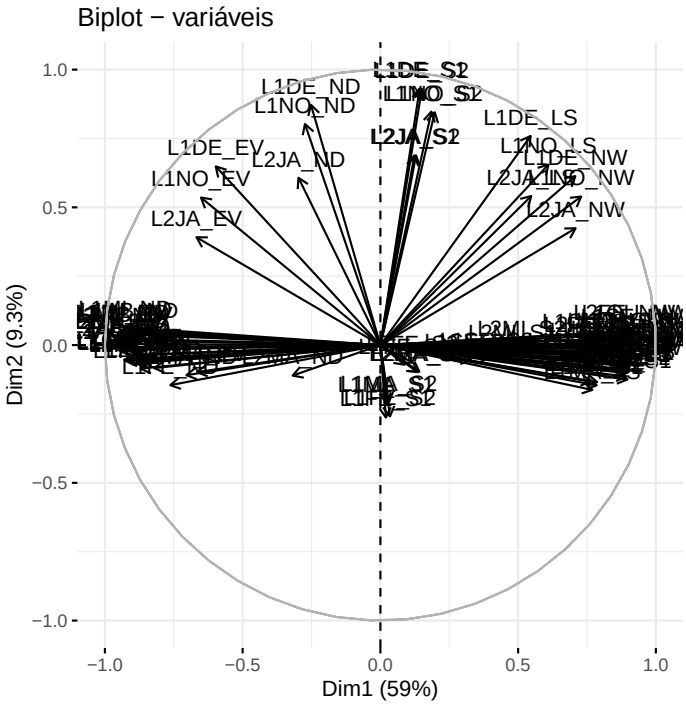
4 Resultados e discussão

As bandas espectrais e os índices de vegetação ao longo de 17 meses totalizaram 102 covariáveis. Dessa forma, aplicamos uma análise de componentes principais com o objetivo de reduzir a dimensionalidade, tornando a análise mais simples. Escolhemos aplicar nos métodos de aprendizado de máquinas as componentes principais que possuem autovalores superiores a 1. A Tabela 1 apresenta a variância explicada pelos 12 componentes principais escolhidos.

Tabela 1: Componentes principais utilizados e porcentagem de variabilidade dos dados explicada.

Componentes principais	Variância (%)	Variância acumulada (%)
1	59,02	59,02
2	9,30	68,32
3	5,27	73,59
4	4,72	78,31
5	4,15	82,46
6	3,87	86,33
7	2,78	89,11
8	2,23	91,34
9	1,90	93,24
10	1,51	94,75
11	1,17	95,92
12	1,00	96,92

Conforme podemos observar, as 12 componentes principais escolhidas explicam juntas 96,92% da variabilidade dos dados. Além disso, as duas primeiras componentes principais foram relacionadas às covariáveis originais dos dados através do gráfico "Biplot"apresentado na Figura 3 a seguir. É possível notar que as 6 covariáveis dos meses de fevereiro a outubro estão altamente correlacionadas à componente principal 1 (eixo horizontal), enquanto a algumas covariáveis dos meses de novembro, dezembro e janeiro estão altamente correlacionadas à componente principal 2 (eixo vertical). Dessa forma, podemos sugerir chamar a componente principal 1 de "índices da estação seca" e a componente principal 2 de "índices no período das águas".



Fonte: Elaborado pelo autor.

Figura 3: "Biplot"referente às covariáveis originais e às duas primeiras componentes principais.

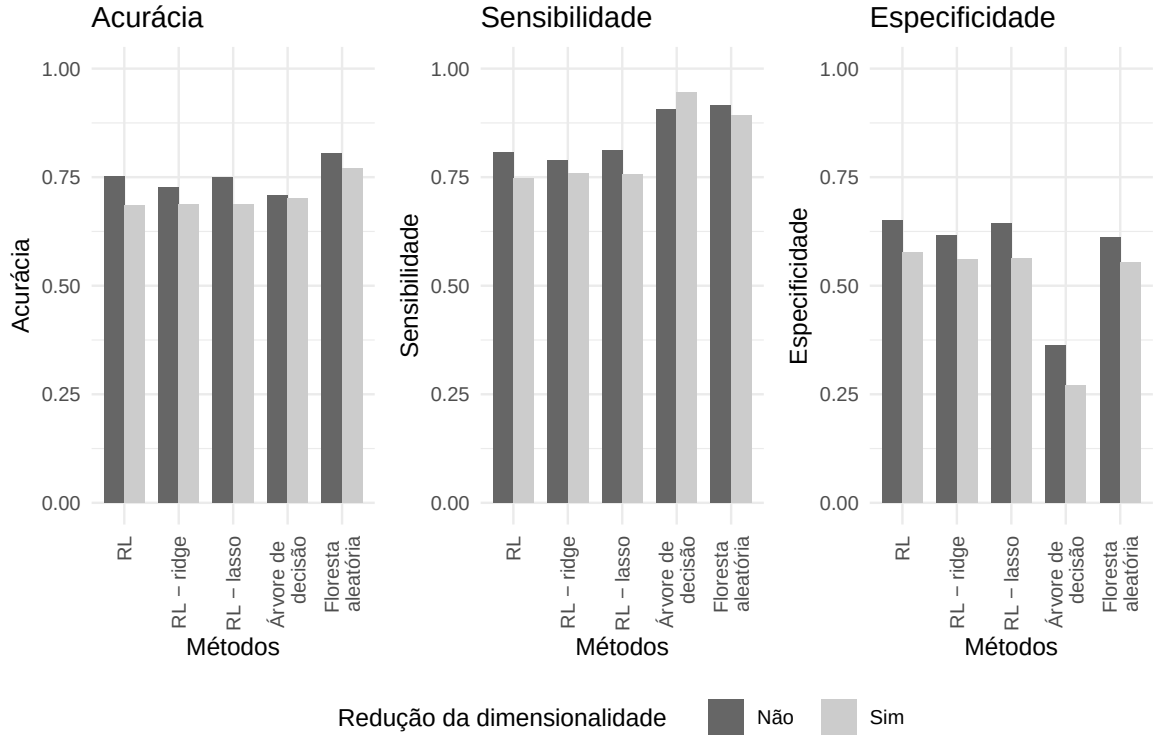
Feita nossa análise, extraímos os valores dos 12 componentes e os juntamos à classificação real dos dados, formando um novo banco de dados. Separamos o novo banco em treino (80%) e teste (20%), da mesma forma que fizemos com o banco inicial e, então, aplicamos esses dados aos métodos de aprendizado de máquina. Na Tabela 2, podemos observar lado a lado as matrizes de confusão dos modelos de predição, respectivamente, da Regressão Logística, das Árvores de Decisão e do Florestas aleatórias com e sem a aplicação do PCA.

Na Figura 4 podemos observar o comparativo da acurácia, da sensibilidade, da especificidade da classificação de cada um dos modelos utilizados. Dos modelos testados o Florestas aleatórias foi o que apresentou maior acurácia, e seus valores de sensibilidade e especificidade não foram muito distantes dos outros modelos, sendo o melhor modelo para ser usado neste contexto. A Regressão Logística fica atrás apenas do Florestas aleatórias nos comparativos apresentados, já a Árvore de Decisão foi o modelo teve o pior desempenho entre os três.

Tabela 2: Matriz de confusão de cada um dos modelos, com e sem aplicação do PCA.

Sem PCA	RL		RL R		RL L		AD		RF	
Referência	0	1	0	1	0	1	0	1	0	1
0	2162	1115	2047	1230	2139	1101	1207	546	2034	488
1	1161	4702	1276	4587	1184	4716	2116	5271	1289	5329
Com PCA	**RL**		**RL R**		**RL L**		**AD**		**RF**	
Referência	0	1	0	1	0	1	0	1	0	1
0	1916	1465	1860	1401	1874	1411	902	317	1840	619
1	1407	4352	1463	4416	1449	4406	2421	5500	1483	5198

RL: regressão logística; RL R: regressão logística com ridge; RL L: regressão logística com lasso; AD: árvore de decisão; RF: Florestas aleatórias.



Fonte: Elaborado pelo autor.

Figura 4: Medidas de desempenho dos métodos na classificação de cana-de-açúcar.

O banco de dados utilizado inicialmente tem 102 covariáveis para cada polígono da área de estudo, e somente na área 6 temos cerca de 46.000 polígonos diferentes. Dessa forma, não é improvável dizer que o tempo para predição de áreas maiores teria um tempo de processamento muito longo. Pensando nisso, analisamos também o tempo necessário para processar os modelos antes e depois da redução de dimensionalidade, que pode ser visto na Tabela 3.

Tabela 3: Tempo de processamento de cada um dos modelos com e sem aplicação do PCA.

Modelos	Sem PCA	Com PCA	Redução
	Tempo (s)	Tempo (s)	(%)
Regressão Logística	7,70	0,03	99,61
Florestas aleatórias	464,29	50,46	89,13
Árvores de Decisão	9,11	1,02	88,80

A seguir, apresentamos a Figura 5 contendo o mapa da região de estudo, comparando a predição da classificação dos segmentos da imagem Landsat pelos três métodos, após PCA, com a classificação real.

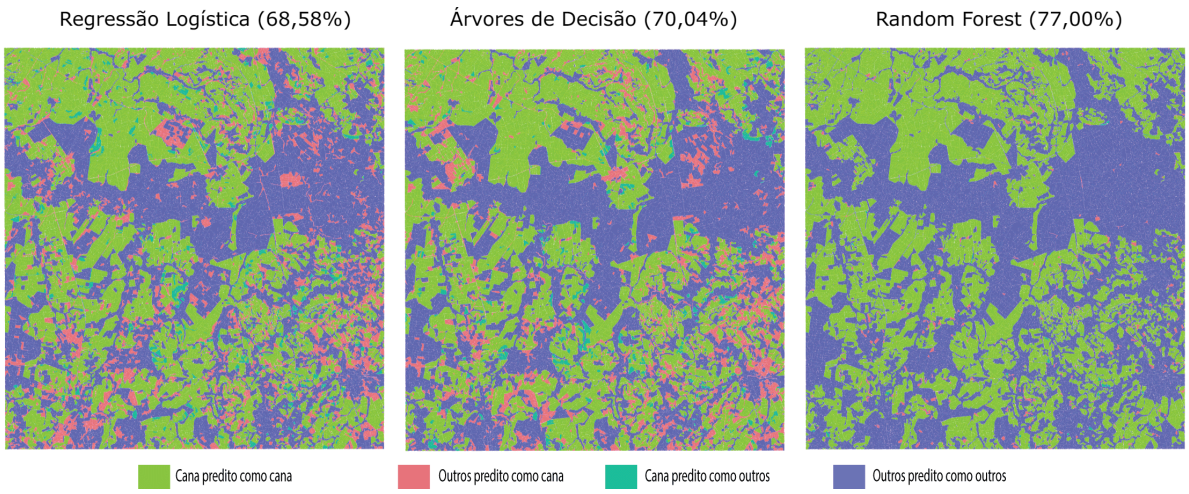


Figura 5: Localização da região de estudo no interior do estado de São Paulo, com a representação dos segmentos do mapa com cores falsas para demonstrar o comparativo dos três modelos de predição utilizados: Regressão Logística, Árvores de Decisão e Florestas aleatórias.

Ao observar a Figura 5, vemos as áreas destacadas nas cores verde e roxo que repre-

sentam as predições corretas do modelo, já as áreas nas cores azul e rosa representam as áreas onde o modelo errou a predição da classificação do uso da terra. Para formar uma imagem completa da área de estudo foi predito além do conjunto de teste também o conjunto de treino. A predição do conjunto de treino só foi usada para formação da imagem, mas não foi considerada na acurácia do modelo. Como o conjunto de treino tem uma acurácia maior, as imagens podem aparentar uma acurácia maior que a descrita. Como os modelos de Regressão Logística penalizada forneceram resultados muito próximos, optamos pela utilização apenas da Regressão Logística sem penalização para elaboração dos comparativos.

Neste trabalho, exploramos alguns algoritmos de aprendizado de máquina para contribuir na automação do mapeamento da cana-de-açúcar. Mensuramos a qualidade dos resultados obtidos pelos diferentes algoritmos a partir de algumas medidas de desempenho de classificação: a acurácia, a sensibilidade e a especificidade. Embora tenhamos alcançado valores satisfatórios para a acurácia e a sensibilidade, notamos valores inferiores para a especificidade nos três métodos, conforme resultados apresentados na Figura 4.

Destacamos que modelos que apresentam baixa especificidade são mais propensos a fornecer resultados falsos positivos, o que significa em nosso problema classificar segmentos como cana-de-açúcar, sendo que na realidade não são. Como os três métodos forneceram baixa especificidade, entendemos que o conjunto de dados considerado no estudo possui características específicas que dificultam a identificação de parte dos segmentos que não são cana-de-açúcar.

Quanto a redução do tempo de processamento pelo modelo, todos obtiveram uma redução maior que 80%. Consideramos que a pequena queda nos parâmetros de acurácia, sensibilidade e especificidade do modelo são consequências aceitáveis para o ganho de agilidade promovidos pela redução de dimensionalidade.

Acreditamos que este trabalho irá contribuir na escolha dos métodos de classificação

da terra e auxiliar na tomada de decisões no cultivo da cana-de-açúcar e atividades correlacionadas através da aplicação realizada, que ilustra o desempenho dos métodos utilizados em casos reais. Esperamos também incentivar o uso de métodos de aprendizado de máquina na área de Ciências Agrárias em outros trabalhos de classificação da terra com diferentes cultivos ou até mesmo na identificação via satélite de pragas em campo.

5 Conclusões

Em relação à proposta científica deste trabalho, realizamos um sólido estudo acerca dos algoritmos de aprendizado de máquina aplicados na automação do mapeamento da cana-de-açúcar. A qualidade dos resultados obtidos pelos diferentes algoritmos foi mensurada e comparada com a literatura através de critérios computacionais e estatísticos. Especificamente, os algoritmos utilizados forneceram acurácia acima de 68%, o que julgamos estar dentro do esperado. Como já discutido na Seção 4, uma limitação encontrada na metodologia aplicada é o número elevado de predição de falsos positivos na classificação dos segmentos. Para buscar reduzir a quantidade de segmentos classificados como falsos positivos, sugerimos como trabalho futuro a implementação de um método de pós-classificação que considere a classificação dos segmentos vizinhos para confirmar a predição de cada segmento e a separação dos índices obtidos durante a período das águas e durante a época seca. Para tornar o processo mais ágil, a análise de componentes principais (PCA) foi considerada uma ferramenta adequada, reduzindo mais de 80% do tempo de processamento dos modelos, porém com uma pequena queda dos parâmetros de acurácia, sensibilidade e especificidade.

Acerca da formação de recursos humanos, este trabalho proporcionou meios para o desenvolvimento científico e investigativo da aluna, estimulando o seu potencial através de novos conhecimentos teóricos e práticos nas áreas de estatística e aprendizado de máquina.

6 Referências Bibliográficas

ARAÚJO, E. D. S. e Santos, J. A. P. (2013). O desenvolvimento da cultura da cana-de-açúcar no brasil e sua relevância na economia nacional. *FACIDER-Revista Científica*, 4(4).

Baatz, M. (2000). Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. *Angewandte geographische informationsverarbeitung*.

Becker, R. S., Alonço, A., Francetto, T., Rodrigues, H. E., Bock, R., e Mendonça, M. (2021). Inovações tecnológicas em máquinas agrícolas para controle de plantas daninhas. *Tecno-Lógica*, 25:98–108.

Bingham, N. H. e Fry, J. M. (2010). *Regression: Linear Models in Statistics*, volume 1. Springer London.

Burkov, A. (2019). *The Hundred Page Machine Learning Book*.

CONAB, C. N. d. A. (2020). Acompanhamento da safra brasileira da cana-de-açúcar, safra 2020/2021. primeiro levantamento, maio de 2020.

Cutts, A. e Graser, A. (2018). *Learn QGIS*. Packt, 4 edition.

da Costa Araújo Filho, M., Meneses, P. R., e Sano, E. E. (2007). Sistema de classificação de uso e cobertura da terra com base na análise de imagens de satélite. *Rev. Bras. Cartogr*, 59.

Dainese, R. C. (2001). Sensoriamento remoto e geoprocessamento aplicado ao estudo temporal do uso da terra e na comparação entre classificação não supervisionada e análise visual.

Delgado, R. C., Sediyama, G. C., Costa, M. H., Soares, V. P., e Andrade, R. G. (2012). Classificação espectral de área plantada com a cultura da cana-de-açúcar por meio da árvore de decisão. *Engenharia Agrícola*, 32:369–380.

Gilbertson, J. K. e van Niekerk, A. (2017). Value of dimensionality reduction for crop differentiation with multi-temporal imagery and machine learning. *Computers and Electronics in Agriculture*, 142:50–58.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Henderson, D., Williams, C. J., e Miller, J. S. (2007). Forecasting late blight in potato crops of southern idaho using logistic regression analysis. *Plant disease*, 91(8):951–956.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*.

Katarya, R., Raturi, A., Mehndiratta, A., e Thapper, A. (2020). Impact of machine learning techniques in precision agriculture. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pages 1–6. IEEE.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Lad, A. M., Bharathi, K. M., Saravanan, B. A., e Karthik, R. (2022). Factors affecting agriculture and estimation of crop yield using supervised learning algorithms. *Materials Today: Proceedings*, 62:4629–4634.

Lê, S., Josse, J., e Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Lillesand, T. M., Kiefer, R. W., e Chipman, J. W. (2015). *Remote sensing and image interpretation*, volume 7.

Luciano, A., Picoli, M., Rocha, J., Franco, H., Sanches, G., Leal, M., e le Maire, G. (2018). Generalized space-time classifiers for monitoring sugarcane areas in brazil. *Remote Sensing of Environment*.

Luciano, A. C. S., Duft, D. G., Picoli, M. C. A., Rocha, J. V., e Le Maire, G. (2019). Estimativa da produtividade de cana-de-açúcar utilizando imagens landsat e random forest. In *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*.

Manly, B. F. J. e Alberto, J. A. N. (2008). *Métodos estatísticos multivariados: uma introdução*. Bookman Editora.

Meira, C. A., Rodrigues, L. H., e Moraes, S. A. (2008). Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. *Tropical Plant Pathology*, 33:114–124.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Morettin, P. e Bussab, W. (2017). *Estatística básica*.

Morettin, P. e Singer, J. (2022). *Estatística e Ciéncia de Dados*.

Nocelli, R. C. F., Zambon, V., Guilherme, O., da Silva, M., e de Castro Morini, M. S. (2017). Histórico da cana-de-açúcar no brasil: contribuições e importância econômica. *Cana-de-açúcar e seus impactos: uma visão acadêmica*, page 13.

Nonato, R. T. e Oliveira, S. R. d. M. (2013). Técnicas de mineração de dados para identificação de Áreas com cana-de-açúcar em imagens Landsat 5. *Engenharia Agrícola*, 33:1268 – 1280.

OECD-FAO (2018). Oecd-fao agricultural outlook 2018-2027.

Oliveira, H. F. d. (2010). Avaliação de modelos de estimativa de produtividade da cana-de-açúcar irrigada em jaíba-mg.

Peres, J. R. R., Junior, E. d. F., e Gazzoni, D. L. (2005). Biocombustíveis uma oportunidade para o agronegócio brasileiro. *Revista de Política Agrícola*, 14(1):31–41.

Prasad, N. R., R., P., e Danodia, A. (2020). Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*.

Rani, S., Kumar, S., T, V. S., Jain, A., Swathi, A., e M, R. K. M. V. N. (2022). Commodities price prediction using various ml techniques. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 277–282.

Rodrigues, L. D. (2010). A cana-de-açúcar como matéria-prima para a produção de biocombustíveis: impactos ambientais e o zoneamento agroecológico como ferramenta para mitigação. *Juiz de Fora-MG, UFJF*.

Ruiz Hidalgo, D., Bacca Cortés, B., e Caicedo Bravo, E. (2021). Dimensionality reduction of hyperspectral images of vegetation and crops based on self-organized maps. *Information Processing in Agriculture*, 8(2):310–327.

Shahwan, T. e Odening, M. (2007). *Forecasting Agricultural Commodity Prices using Hybrid Neural Networks*, pages 63–74. Springer Berlin Heidelberg, Berlin, Heidelberg.

Silva, J. M. e Konradt-Moraes, L. C. (2012). Vantagens e desvantagens dos biocombustíveis e dos combustíveis fósseis. *Anais do SEMEX*, (5).

Souza, Z., Cerri, D., Colet, M., Rodrigues, L. H., Graziano Magalhães, P., e Mandoni, R. (2010). Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. *Ciência Rural*, 40.

Tay, J. K., Narasimhan, B., e Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31.

Therneau, T. M., Atkinson, E. J., e Foundation, M. (2022). An introduction to recursive partitioning using the rpart routines.

Van Der Maaten, L., Postma, E., e Van den Herik, J. (2009). Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10(66-71):13.

Vieira, M., Formaggio, A., Rennó, C., Atzberger, C., Aguiar, D., e Mello, M. (2012). Object based image analysis and data mining applied to a remotely sensed landsat time-series to map sugarcane over large areas. *Remote Sensing of Environment*, 123:553–562.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zhai, Y., Wang, N., Zhang, L., Hao, L., e Hao, C. (2020). Automatic crop classification in northeastern china by improved nonlinear dimensionality reduction for satellite image time series. *Remote Sensing*, 12(17).