

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luiz Felipe Cavalheiro dos Santos

**ANÁLISE DO IMPACTO DAS MUDANÇAS CLIMÁTICAS NA
AGRICULTURA DA REGIÃO GEOGRÁFICA INTERMEDIÁRIA DE
SANTA MARIA UTILIZANDO TÉCNICAS DE MACHINE LEARNING**

Santa Maria, RS
2025

Luiz Felipe Cavaleiro dos Santos

**ANÁLISE DO IMPACTO DAS MUDANÇAS CLIMÁTICAS NA AGRICULTURA DA
REGIÃO GEOGRÁFICA INTERMEDIÁRIA DE SANTA MARIA UTILIZANDO TÉCNICAS
DE MACHINE LEARNING**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação**.

Orientador: Prof. Joaquim Vinicius Carvalho Assunção

Santa Maria, RS
2025

Luiz Felipe Cavalheiro dos Santos

**ANÁLISE DO IMPACTO DAS MUDANÇAS CLIMÁTICAS NA AGRICULTURA DA
REGIÃO GEOGRÁFICA INTERMEDIÁRIA DE SANTA MARIA UTILIZANDO TÉCNICAS
DE MACHINE LEARNING**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação**.

Aprovado em 3 de julho de 2025:

**Joaquim Vinicius Carvalho Assunção, Dr. (UFSM)
(Presidente/Orientador)**

Daniel Welfer, Dr. (UFSM)

Antônio do Nascimento Leães Neto, Me. (Uniritter)

Santa Maria, RS
2025

DEDICATÓRIA

À minha família, pois foi ela que me deu a oportunidade de estudar, de sair de casa rumo à uma cidade muito maior e de ingressar em uma das grandes universidades do país. Lembro que me inscrevi no SISU por curiosidade, apenas para descobrir até onde eu conseguiria chegar. Fiz isso em silêncio, sem contar a ninguém e só revelei quando fui aprovado. A partir desse momento, fizeram de tudo para que eu pudesse estar aqui. Mesmo enfrentando muitas dificuldades, minha família me ofereceu todo o suporte: apoio emocional, ajuda financeira, comida enviada com carinho e, acima de tudo, uma confiança no meu potencial. Quando pensei em desistir, lembrar de todo o esforço que já haviam feito por mim foi uma das coisas que me deu forças para continuar. Por isso, esta conquista definitivamente não é apenas minha. É de toda a minha família. É o reflexo de cada sacrifício, de cada gesto de amor e de cada batalha vencida juntos.

AGRADECIMENTOS

Agradeço, ao meu orientador, por toda a atenção, apoio e confiança. Sempre estive disponível para ouvir minhas ideias, me orientar com sugestões valiosas e, acima de tudo, respeitar meu tempo ao longo do processo. Sua compreensão e confiança na minha capacidade foram fundamentais para que este trabalho se concretizasse. Estendo meu agradecimento a todas as pessoas que fizeram parte da minha trajetória desde que cheguei a Santa Maria. Seria impossível citar todas e ao tentar, correria o risco de ser injusto. Agradeço aos meus familiares, que sempre estiveram comigo nos momentos mais difíceis; meus amigos, que me ofereceram caronas, conselhos e companhia quando mais precisei; e à quem conheci durante os anos de graduação, com quem vivi momentos especiais e inesquecíveis. Cada gesto, por menor que pareça, foi essencial para que eu chegasse até aqui. À todos que de alguma forma, caminharam comigo nessa jornada: meu sincero e eterno obrigado.

RESUMO

ANÁLISE DO IMPACTO DAS MUDANÇAS CLIMÁTICAS NA AGRICULTURA DA REGIÃO GEOGRÁFICA INTERMEDIÁRIA DE SANTA MARIA UTILIZANDO TÉCNICAS DE MACHINE LEARNING

AUTOR: Luiz Felipe Cavalheiro dos Santos
Orientador: Joaquim Vinicius Carvalho Assunção

Este trabalho tem como objetivo analisar o impacto das mudanças climáticas na produção de soja e arroz na Região Geográfica Intermediária de Santa Maria (RS), utilizando técnicas de Machine Learning. Para isso, serão empregados dados históricos do clima (INMET) e produção agrícola (IBGE) durante a série temporal de 1994 a 2023 em relação aos 40 municípios da região. Inicialmente, se aplica os algoritmos Random Forest e XGBoost para identificar e ranquear a importância das variáveis climáticas na produção agrícola. Em seguida, se utiliza a arquitetura LSTM (*Long Short-Term Memory*) para prever a produção de soja e arroz para as próximas 10 safras (2025-2034), considerando 4 cenários climáticos distintos do modelo CMIP6. Os resultados são apresentados em mapas, visando facilitar a compreensão e permitindo uma visualização anual para cada cenário e cultivo, possibilitando uma comparação entre a estimativa da produção dos municípios da região, bem como, do impacto causado pelas mudanças climática em cada um dos 40 municípios.

Palavras-chave: LSTM; Random Forest; XGBoost; Mudanças Climáticas; Soja; Arroz.

ABSTRACT

ANALYSIS OF THE IMPACT OF CLIMATE CHANGES ON AGRICULTURE IN THE INTERMEDIATE GEOGRAPHIC REGION OF SANTA MARIA USING MACHINE LEARNING TECHNIQUES

AUTHOR: Luiz Felipe Cavaleiro dos Santos
ADVISOR: Joaquim Vinicius Carvalho Assunção

This study aims to analyze the impact of climate changes on soybean and rice production in the Intermediate Geographic Region of Santa Maria (RS), using Machine Learning techniques. For this purpose, historical climate data (INMET) and agricultural production data (IBGE) from 1994 to 2023 for the 40 municipalities in the region will be used. Initially, the Random Forest and XGBoost algorithms will be applied to identify and rank the importance of climatic variables in agricultural production. Subsequently, the LSTM (Long Short-Term Memory) architecture will be employed to forecast soybean and rice production for the next 10 harvests (2025-2034), considering four distinct climate scenarios from the CMIP6 model. The results will be presented through maps, aiming to facilitate understanding and allow annual visualization for each scenario and crop, enabling a comparison of production estimates among the municipalities as well as the impacts of climate change on each of the 40 municipalities.

Keywords: LSTM; Random Forest; XGBoost; Climate Changes; Soybean; Rice.

LISTA DE FIGURAS

Figura 1 – Regiões Geográficas Intermediárias do RS	16
Figura 2 – Regiões Geográficas Imediatas contidas na Região Intermediária de SM ..	16
Figura 3 – Calendário do cultivo de soja no RS para a safra 2022/2023	17
Figura 4 – Fenologia da Soja.	18
Figura 5 – Fenologia do Arroz.	19
Figura 6 – Etapas do processo KDD.	20
Figura 7 – Esquema do Random Forest.	23
Figura 8 – Arquitetura de uma célula no modelo LSTM.	25
Figura 9 – Produção de Arroz no RS em 2023.	32
Figura 10 – Produção de Soja no RS em 2023.	32
Figura 11 – Gráficos de Dispersão do conjunto de testes para Quantidade Produzida nas Séries de 30 e 20 anos e Rendimento Médio na Série de 20 anos, respectivamente.	52
Figura 12 – Gráficos de Dispersão do conjunto de testes para Rendimento Médio de Soja nas Séries de 30 e 20 anos, respectivamente.	53
Figura 13 – Gráficos de Dispersão do conjunto de testes para Quantidade Produzida de Soja nas Séries de 30 (Modelo LSTM1) e 20 anos (Modelo LSTM2), respectivamente.	54
Figura 14 – Histórico de Quantidade Produzida de Arroz.	60
Figura 15 – Histórico de Rendimento Médio do Arroz.	60
Figura 16 – Previsão Quantidade Produzida de Arroz com Série de 30 anos.	61
Figura 17 – Previsão Rendimento Médio do Arroz com Série de 30 anos.	62
Figura 18 – Previsão Quantidade Produzida de Arroz com Série de 20 anos.	63
Figura 19 – Histórico de Quantidade Produzida de Soja.	64
Figura 20 – Histórico de Rendimento Médio da Soja.	64
Figura 21 – Previsão Quantidade Produzida de Soja com Série de 30 anos.	65
Figura 22 – Previsão Rendimento Médio da Soja com Série de 30 anos.	66
Figura 23 – Previsão Quantidade Produzida de Soja com Série de 20 anos.	67
Figura 24 – Previsão Rendimento Médio da Soja com Série de 20 anos.	68
Figura 25 – Precipitação média nos meses de safra.	72
Figura 26 – Temperatura média nos meses de safra.	72
Figura 27 – Radiação solar nos meses de safra.	73
Figura 28 – Umidade do solo nos meses de safra.	73

LISTA DE TABELAS

TABELA 1 – Unidades de medida das variáveis climáticas utilizadas	39
TABELA 2 – Comparação entre os modelos Random Forest e XGBoost.....	42
TABELA 3 – Comparação entre os modelos LSTM1, LSTM2 e LSTM3	45
TABELA 4 – Desempenho dos Modelos Random Forest para Rendimento Médio e Quantidade Produzida de Arroz	47
TABELA 5 – Desempenho dos Modelos Random Forest para Rendimento Médio e Quantidade Produzida de Soja	48
TABELA 6 – Desempenho dos Modelos XGBoost para Rendimento Médio e Quantidade Produzida de Arroz	49
TABELA 7 – Desempenho dos Modelos XGBoost para Rendimento Médio e Quantidade Produzida de Soja.....	50
TABELA 8 – Desempenho dos Modelos LSTM para Rendimento Médio e Quantidade Produzida de Arroz.....	51
TABELA 9 – Desempenho dos Modelos LSTM para Rendimento Médio e Quantidade Produzida de Soja.....	53
TABELA 10 – Importância das Variáveis Climáticas nos Dados Históricos.....	56
TABELA 11 – Importância das Variáveis Climáticas para o Cultivo de Arroz	57
TABELA 12 – Importância das Variáveis Climáticas para o Cultivo de Soja	58

LISTA DE QUADROS

Quadro 1 – Variáveis selecionadas no sistema BDMEP - INMET.	35
Quadro 2 – Ordem de aplicação dos filtros no sistema do ESGF.	37

LISTA DE ABREVIATURAS

BBC	British Broadcasting Corporation
IRGA	Instituto Rio Grandense do Arroz
RS	Rio Grande do Sul
EMATER/RS	Empresa de Assistência Técnica e Extensão Rural do Rio Grande do Sul
ASCAR	Associação Sulina de Crédito e Assistência Rural
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
SIDRA	Sistema IBGE de Recuperação Automática
CMIP6	Coupled Model Intercomparison Project Phase 6
WCRP	World Climate Research Programme
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
MAPA	Ministério da Agricultura e Pecuária
SEAPI	Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
KDD	Knowledge Discovery from Data
IA	Inteligência Artificial
OMM	Organização Meteorológica Mundial
ICSU	Conselho Internacional para a Ciência
COI	Comissão Oceanográfica Intergovernamental
SSP	Shared Socioeconomic Pathways
ESGF	Earth System Grid Federation
PCMDI	Program for Climate Model Diagnosis and Intercomparison
VE	Vegetativa de Emergência
VC	Vegetativa de Cotilédone
FAO	Food and Agriculture Organization of the United Nations
RMSE	Root Mean Square Error
R^2	Coeficiente de Determinação
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
ECMWF	European Centre for Medium-Range Weather Forecasts
XGBoost	Extreme Gradient Boosting

SUMÁRIO

1	INTRODUÇÃO	13
1.1	ESTRUTURAÇÃO DO TRABALHO	14
2	REFERENCIAL TEÓRICO	15
2.1	CONTEXTUALIZAÇÃO DA REGIÃO DE ESTUDO	15
2.2	CICLO PRODUTIVO DE SOJA	17
2.3	CICLO PRODUTIVO DO ARROZ	18
2.4	MINERAÇÃO DE DADOS E KDD	19
2.4.1	Seleção	20
2.4.2	Pré-processamento	20
2.4.3	Transformação	20
2.4.4	Mineração de Dados	21
2.4.5	Interpretação	21
2.5	APRENDIZADO DE MÁQUINA	21
2.5.1	Random Forest	22
2.5.2	XGBoost	23
2.5.3	Long Short-Term Memory (LSTM)	24
2.6	MÉTODOS DE AVALIAÇÃO	25
2.6.1	RMSE	25
2.6.2	MAPE	26
2.6.3	R^2	26
2.7	MODELO CLIMÁTICO CMIP6	27
2.7.1	Cenários SSP	27
3	TRABALHOS RELACIONADOS	29
4	METODOLOGIA	31
4.1	JUSTIFICATIVA DA REGIÃO DE ESTUDO	31
4.2	SELEÇÃO E PRÉ-PROCESSAMENTO DE DADOS	33
4.2.1	Dados dos Municípios	33
4.2.2	Dados de Produção Agrícola	33
4.2.3	Dados Climáticos Históricos (INMET e NASA POWER)	34
4.2.4	Dados Climáticos Futuros (CMIP6)	35
4.2.5	Tratamento de Dados Ausentes	37
4.2.6	Associação de Dados Climáticos e Agrícolas aos Municípios	37
4.3	MINERAÇÃO DE DADOS	39
4.3.1	Aplicação de Random Forest e XGBoost	40
4.3.2	Aplicação de LSTM	42
4.4	INTERPRETAÇÃO DOS RESULTADOS	46

5	RESULTADOS	47
5.1	DESEMPENHO DOS MODELOS RANDOM FOREST E XGBOOST	47
5.2	DESEMPENHO DOS MODELOS LSTM.....	51
5.3	IMPORTÂNCIA DAS VARIÁVEIS CLIMÁTICAS	54
5.4	PREVISÕES PARA AS PRÓXIMAS 10 SAFRAS DE SOJA E ARROZ	58
5.4.1	Previsões para o cultivo de Arroz	59
5.4.2	Previsões para o cultivo de Soja	63
5.5	SISTEMA INTERATIVO PARA VISUALIZAÇÃO DOS RESULTADOS	68
6	CONCLUSÃO	70
6.1	TRABALHOS FUTUROS	71
	APÊNDICE A – GRÁFICOS COMPLEMENTARES	72
	REFERÊNCIAS BIBLIOGRÁFICAS	74

1 INTRODUÇÃO

Nos últimos anos o estado do Rio Grande do Sul tem enfrentado diversos fenômenos climáticos extremos como *La Niña* e *El Niño*, estiagens, ondas de calor, tornados e enchentes (BBC News Brasil, 2024), entre os diversos problemas causados por esses fenômenos, um dos principais setores da economia do estado, a agricultura, tem sofrido perdas consideráveis. O arroz e a soja são culturas fundamentais para a economia gaúcha e mais especificamente da Região Geográfica Intermediária de Santa Maria (RS), composta por 40 municípios, compreender como o clima afeta sua produtividade é crucial para planejamento e mitigação de riscos. Segundo EMATER/RS-ASCAR (2024), a região central do Rio Grande do Sul, além de ser a mais atingida pelas enchentes de 2024, conta com o maior número de pequenos produtores, gerando impacto econômico muito grande nessa região.

O Valor Bruto da Produção Agropecuária gaúcha no ano de 2023 foi de R\$ 98,2 bilhões, sendo a soja com 31% e o arroz com 13%, os produtos de maior relevância no setor agropecuário, e que somados representam 44% do valor total, segundo dados apresentados pela Secretaria de Agricultura do estado do Rio Grande do Sul em 2024, referentes à safra 2023/2024. Os 40 municípios da Região Geográfica Intermediária de Santa Maria se destacam como grandes produtores de soja, sendo que muitos deles também se destacam no cultivo de arroz. Vale ressaltar que na safra de 2023/2024 o Rio Grande do Sul foi o 3º maior produtor de soja do Brasil, colhendo 18,3 milhões de toneladas, e ainda, o maior produtor de arroz no país, sendo responsável por 68% da produção nacional.

Considerando a relevância do tema, estudos sobre o impacto das mudanças climáticas na agricultura são necessários para dar base científica às autoridades e órgãos competentes para que planejem ações com o objetivo de conter danos na agricultura em caso de novos desastres naturais no futuro. Dessa forma, se justifica a escolha do tema deste trabalho, o estudo tem como objetivo analisar o impacto das mudanças climáticas na produção de soja e arroz na Região Geográfica Intermediária de Santa Maria (RS) utilizando técnicas de *Machine Learning*. São utilizados dados históricos fornecidos pelo INMET e NASA POWER referentes ao clima dos 40 municípios que compõem a região e também dados históricos fornecidos pelo IBGE referentes à produção agrícola de soja e arroz destes municípios.

O objetivo é desenvolver modelos preditivos capazes de estimar a produção de soja e arroz na região para os próximos 10 anos (2025-2034), com base em diferentes cenários climáticos. Para isso, combinam-se técnicas de mineração de dados com aprendizado de máquina, para encontrar relações entre os dados climáticos e de produção agrícola, bem como a relevância de cada variável. Sendo assim, foram escolhidos os algoritmos *Random Forest* e *XGBoost*, pois possuem características adequadas para encontrar e ranquear a

importância de cada variável na relação (BREIMAN, 1996). Para a previsão futura da produção agrícola a abordagem utiliza LSTM (*Long Short-Term Memory*), uma arquitetura de rede neural recorrente (RNN) que se destaca no processamento de séries temporais (HOCHREITER; SCHMIDHUBER, 1997), e portanto, será o algoritmo utilizado nesta fase.

Para isso, são utilizados dados históricos da produção agrícola dos cultivos de soja e arroz de uma série temporal, isto é, de um período de 30 anos (1994-2023) obtidos através do sistema SIDRA do IBGE, além de dados climáticos também representados por uma série temporal no mesmo período de tempo, são obtidos por meio do sistema do Banco de Dados Meteorológico para Ensino e Pesquisa, pertencente ao INMET que contém dados climáticos de estações meteorológicas.

Serão abordados 4 cenários climáticos distintos, fornecidos pelo modelo CMIP6 (*Coupled Model Intercomparison Project - Phase 6*), que é um projeto internacional liderado pela WCRP (*World Climate Research Programme*) que reúne centros de pesquisa climática do mundo todo para comparar e aperfeiçoar modelos climáticos globais e tem como objetivo simular o clima do planeta Terra com diferentes cenários de emissões de gases. O modelo preditivo será gerado para cada um desses cenários. Os dados históricos do modelo CMIP6 também são utilizados como complemento aos dados das estações meteorológicas do INMET, já que as mesmas não possuem variáveis climáticas relevantes para o cultivo dos produtos que são estudados.

Os resultados da pesquisa são apresentados em mapas para facilitar a compreensão dos resultados obtidos, permitindo uma visualização anual para cada cenário e cultivo, além de permitir uma comparação entre a estimativa da produção dos municípios da região, bem como, do impacto causado pelas mudanças climática em cada um dos 40 municípios.

1.1 ESTRUTURAÇÃO DO TRABALHO

Este trabalho está estruturado da seguinte forma: o capítulo 1 apresenta a introdução ao tema, contextualizando a importância dos cultivos de soja e arroz no estado e na região de interesse, além de expor os objetivos e a justificativa do estudo. Já no capítulo 2, intitulado Referencial Teórico, são abordadas as técnicas computacionais que serão utilizadas para desenvolvimento desta pesquisa. No capítulo 3, são analisados trabalhos relacionados a este tema, fazendo uma breve revisão e apontamentos das respectivas abordagens. O capítulo 4, é destinado à metodologia utilizada no desenvolvimento do trabalho, detalhando a organização desde a concepção da ideia até as estratégias utilizadas para obtenção de conhecimento, definição e organização dos *datasets* além da aplicação das ferramentas computacionais. Os resultados são apresentados no capítulo 5, acompanhados de uma análise crítica sobre seus impactos e significados. Por fim, o capítulo 6 é reservado para a conclusão e apresentação de sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo serão expostos os principais fundamentos teóricos envolvidos na elaboração deste trabalho, fornecendo a base necessária para a compreensão do tema abordado. Busca-se proporcionar uma imersão nos aspectos da pesquisa desenvolvida, tanto no que diz respeito à agricultura quanto aos modelos e técnicas computacionais aplicadas. Os tópicos estão divididos em seções e subseções abrangendo: contextualização da região de estudo, ciclo produtivo de soja e do arroz, mineração de dados, aprendizado de máquina, métodos de avaliação e modelo climático.

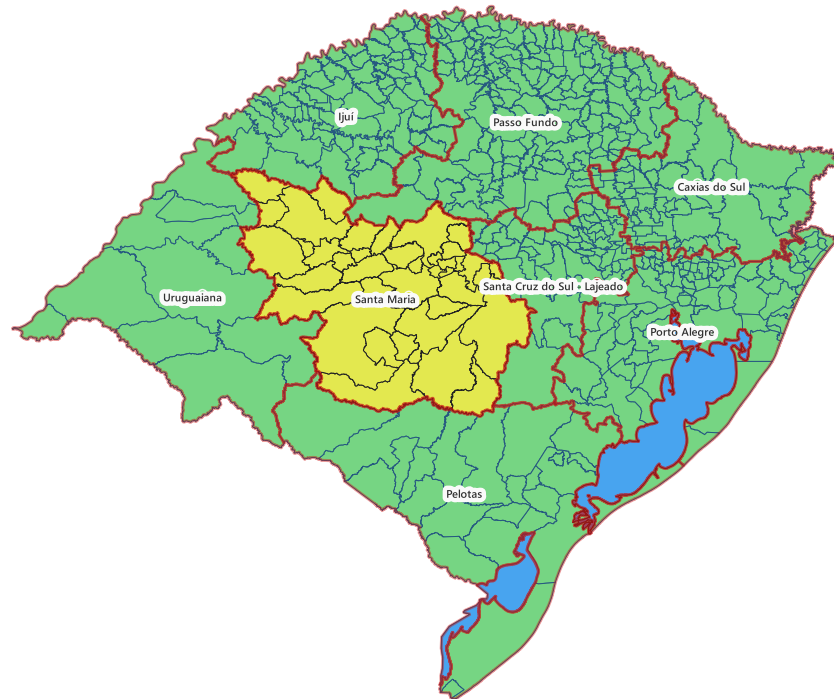
2.1 CONTEXTUALIZAÇÃO DA REGIÃO DE ESTUDO

Criadas no ano de 2017, as regiões geográficas imediatas e intermediárias, visam atualizar o quadro regional do país produzidos na década de 1980 (Instituto Brasileiro de Geografia e Estatística - IBGE, 2017), sucedendo as microrregiões e mesorregiões, respectivamente. A nova regionalização do país reflete as transformações econômicas, demográficas, políticas e ambientais ocorridas nas últimas décadas. Essas novas regiões baseiam-se principalmente na análise da rede urbana e infraestrutura do Brasil, promovendo uma compreensão mais precisa da realidade atual do território nacional. Sendo assim, serve como base para planejamento e tomada de decisões de órgãos públicos e privados. A proposta também oferece subsídios para a pesquisa acadêmica, contribuindo portanto para o conhecimento geográfico do Brasil contemporâneo.

A região geográfica intermediária de Santa Maria, apresentada na Figura 1, é composta por 40 municípios, e abrange as regiões geográficas imediatas de Santa Maria, Santiago, São Gabriel-Caçapava do Sul e Cachoeira do Sul. Todos os municípios dessa região possuem destaque na produção de a soja, além disso, boa parte deles, localizados mais ao sul, se destacam também como grandes produtores de arroz, sendo estes cultivos os dois principais do setor agropecuário, como já mencionado no capítulo anterior.

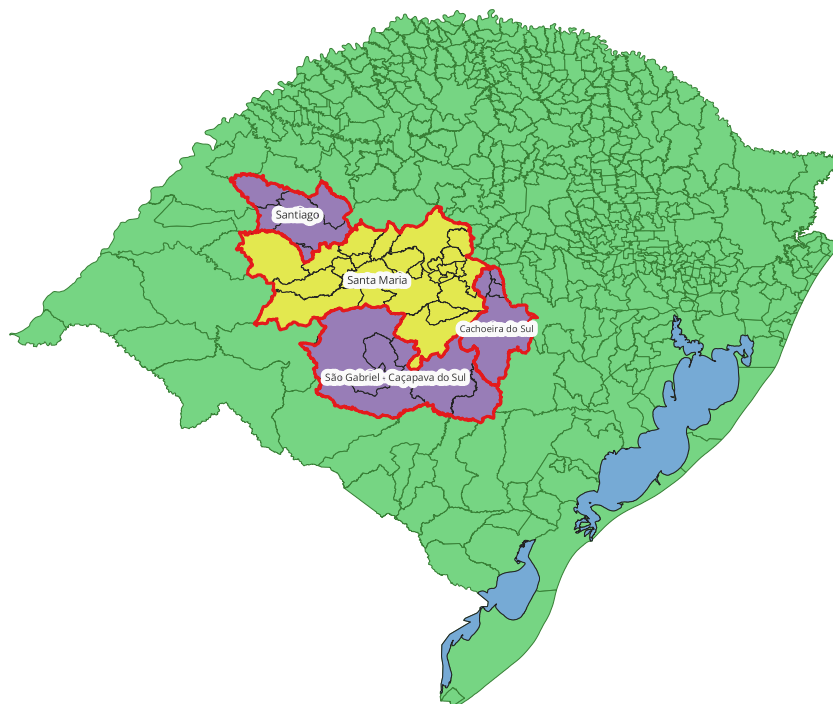
A região geográfica imediata de Santa Maria é composta por 25 municípios, em um comparativo com a região intermediária, pode-se perceber cidades pertencentes a região intermediária e que se destacam na produção de arroz não fazem parte da região imediata. Sendo assim se justifica a escolha pela região geográfica intermediária de Santa Maria pela combinação dos fatores de maior abrangência de municípios e destaque dos mesmos na produção dos principais produtos do setor agropecuário gaúcho.

Figura 1 – Regiões Geográficas Intermediárias do RS



Fonte: Elaborado pelo autor com base em dados do IBGE.

Figura 2 – Regiões Geográficas Imediatas contidas na Região Intermediária de SM

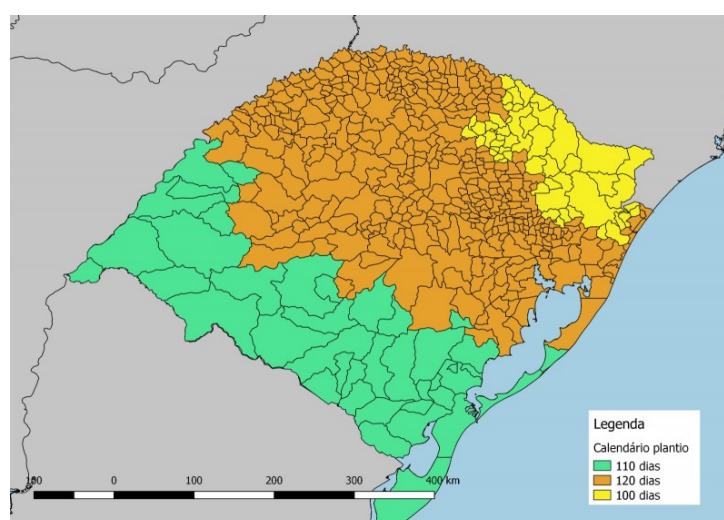


Fonte: Elaborado pelo autor com base em dados do IBGE.

2.2 CICLO PRODUTIVO DE SOJA

O plantio de soja no Rio Grande do Sul tem início entre os meses de outubro e novembro, sendo os melhores resultados obtidos geralmente quando a semeadura é realizada em novembro (BARNI; MATZENAUER, 2014). A colheita é realizada entre os meses de fevereiro e março do ano seguinte, o calendário oficial das safras é definido anualmente pelo Ministério da Agricultura e Pecuária (MAPA). No ano de 2022, a pedido da Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação (Seapi), o calendário de semeadura da soja no RS foi dividido em 3 regiões, com safras durando 100, 110 e 120 dias, conforme mostrado na Figura 3.

Figura 3 – Calendário do cultivo de soja no RS para a safra 2022/2023



Fonte: (Governo do Estado do Rio Grande do Sul, 2024)

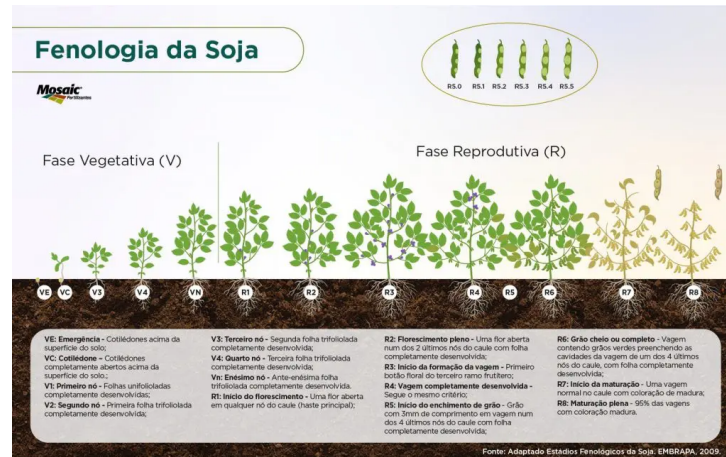
O ciclo produtivo da soja pode ser dividido em duas fases principais segundo FEHR; CAVINESS (1977) vegetativa e reprodutiva. É considerado estágio vegetativo desde o plantio até o início da floração, alguns pontos críticos nessa fase são: temperaturas abaixo de 10°C que podem atrasar o crescimento; chuvas excessivas ou estiagem precoce que podem afetar o estabelecimento da cultura; e a radiação solar que é essencial para estimular o crescimento foliar.

O desenvolvimento da soja inicia-se com as fases Vegetativa de Emergência (VE) e Vegetativa de Cotilédone (VC), na sequência, se inicia o desenvolvimento vegetativo, caracterizado pelas fases de V1 a Vn, nas quais a planta passa a desenvolver folhas trifolioladas em sequência (V1, V2, V3, V4, etc.).

O estágio reprodutivo se estende do florescimento à maturidade fisiológica, o florescimento (R1-R2) inicia com o surgimento das primeiras flores e nele, é definido o número potencial de frutos. Na formação das vagens (R3-R4), nos nós superiores da planta, as vagens são desenvolvidas. Durante o enchimento de grãos (R5-R6), os grãos se desenvolvem até preencher as cavidades das vagens e por fim, na maturação (R7-R8), a

planta atinge a coloração madura, indicando que os grãos estão prontos para a colheita. É durante o ciclo reprodutivo que as condições climáticas possuem maior impacto sobre a planta. Pode-se observar de forma visual na Figura 4, o ciclo produtivo da soja, também conhecido como fenologia.

Figura 4 – Fenologia da Soja.



Fonte: (Nutrição de Safras, 2025)

2.3 CICLO PRODUTIVO DO ARROZ

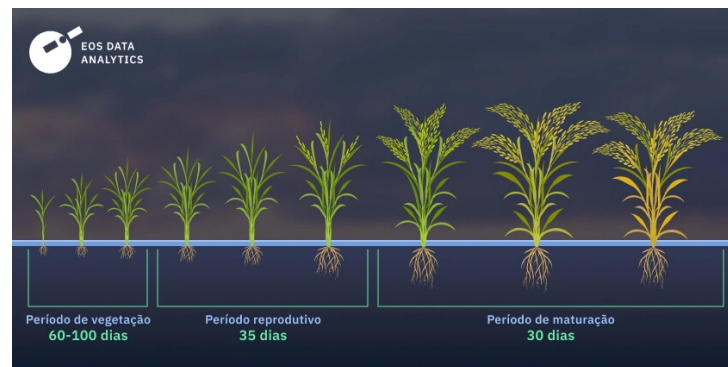
O plantio do arroz no Brasil possui duas modalidades, o **irrigado**, cultivado em áreas de controle hídrico, solo úmido e inundação, comum na região Sul do Brasil, caracteriza-se pela alta produtividade. Já o arroz de **sequeiro** depende exclusivamente da precipitação pluvial, sendo cultivado em áreas de menor umidade, utilizado principalmente no Nordeste do país (Empresa Brasileira de Assistência Técnica e Extensão Rural; Empresa Brasileira de Pesquisa Agropecuária, 1981). No Rio Grande do Sul, as regiões da Depressão Central e da Fronteira Oeste, concentram as maiores áreas alagáveis adaptadas ao cultivo (Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação do Rio Grande do Sul, 2024).

O período de cultivo do arroz, inicia com o plantio entre os meses de outubro e novembro e a colheita ocorre entre fevereiro e março do ano seguinte, o calendário é definido anualmente pelo MAPA de acordo com as condições climáticas. O ciclo do arroz irrigado dura em média de 110 a 140 dias. Segundo COUNCE; L.; FEHR (2000), o ciclo produtivo pode ser dividido em três fases principais: vegetativa (V), reprodutiva (R) e maturação (M).

A fase vegetativa começa quando a planta emerge e vai até o momento no qual a panícula (estrutura responsável pela produção de flores e posteriormente de grãos) começa a se formar, nessa fase a irrigação constante é de grande importância, para o cres-

cimento da planta, temperaturas acima de 20°C são fundamentais. Na fase reprodutiva, que vai da formação da panícula até a floração, é o momento mais delicado, pois a planta é mais sensível a problemas com a água ou com o calor, essas condições adversas diminuem a quantidade de grãos. A fase de maturação, inicia quando os grãos começam a se formar e se estende até maturidade completa, quando a planta está pronta para a colheita, a medida que os grãos amadurecem, a planta necessita cada vez menos de água. O ciclo fenológico do arroz pode ser observado na Figura 5.

Figura 5 – Fenologia do Arroz.

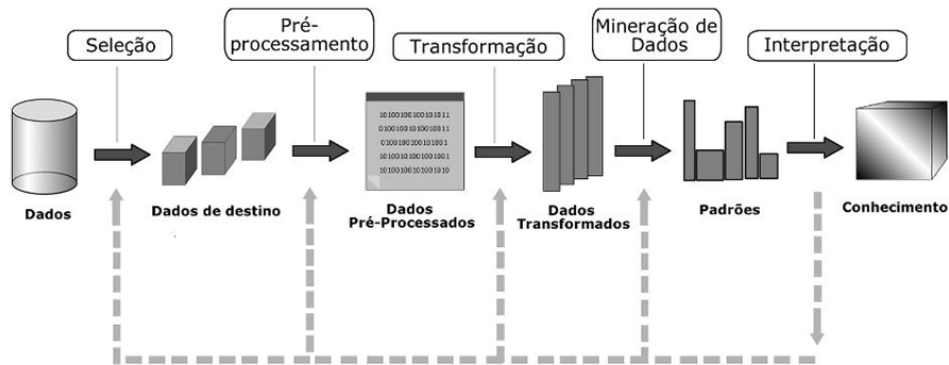


Fonte: (EOS Data Analytics, 2024)

2.4 MINERAÇÃO DE DADOS E KDD

Atualmente, a maioria das operações e atividades que realizamos no cotidiano são feitas com auxílio da tecnologia, sendo registradas computacionalmente, desse modo, muitas informações vão se acumulando em grandes bases de dados. O termo mineração de dados surgiu inicialmente como um sinônimo de KDD (*Knowledge Discovery from Data*), traduzindo para o português, Descoberta de Conhecimento em Bases de Dados, porém, é apenas uma das etapas do processo KDD que por sua vez, utiliza conceitos de base de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial, e foi dividido por FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996) nas seguintes etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação. O objetivo principal do KDD é transformar dados em informações, informações em conhecimento e conhecimento em ação. Neste trabalho, será utilizado o processo KDD como estrutura metodológica para organizar o fluxo de análise dos dados climáticos e de produção agrícola, garantindo uma abordagem sistemática e eficiente para extração de conhecimento a partir dos dados disponíveis.

Figura 6 – Etapas do processo KDD.



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) (Traduzido).

2.4.1 Seleção

A primeira etapa do processo é a seleção de dados, nela, os dados relevantes para análise são selecionados, isso inclui a identificação de fontes confiáveis, definições de critérios de inclusão e exclusão de dados. Ao final dessa etapa, deve-se obter *datasets* potencialmente relevantes.

2.4.2 Pré-processamento

Após a seleção dos dados, a próxima etapa é o pré-processamento, no qual os dados brutos obtidos no passo anterior são preparados para análise, isso inclui 3 procedimentos: **limpeza de dados** que consiste em remover ruídos e dados inconsistentes, além do tratamento de valores ausentes e normalização dos dados caso necessário; a **integração de dados** consiste em integrar múltiplas fontes de dados para serem combinadas em um arquivo; por fim, segundo HAN; KAMBER; PEI (2011), a **seleção** diz respeito a escolha dos dados propícios a serem analisados.

2.4.3 Transformação

Nesta etapa, os dados são transformados nos formatos apropriados para a mineração, isso pode envolver técnicas de agregação, redução de dimensionalidade e codificação de dados, este processo depende do tipo de algoritmo que será utilizado.

2.4.4 Mineração de Dados

Esta etapa pode ser considerada o cerne do processo KDD, é nela que são aplicados algoritmos inteligentes de mineração para descobrir padrões, relações e *insights* nos dados que previamente eram desconhecidos. Existem diferentes técnicas de algoritmos de mineração de dados, as 4 principais de acordo com FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996), são: **associação** (consiste em identificar padrões frequentes e relações entre atributos em grandes volumes de dados), **classificação** (consiste em aprender uma função que atribui itens de dados a classes predefinidas), **clusterização** (identifica grupos naturais nos dados sem rótulos prévios), e **regressão** (busca prever um valor numérico contínuo com base em variáveis de entrada).

2.4.5 Interpretação

Após a descoberta de padrões, é de grande relevância saber como apresentar o conhecimento extraído, ou ao menos expor os resultados de forma clara e acessível, para que outras pessoas possam compreender, tirando conclusões.

2.5 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (*Machine Learning*) deu seus primeiros passos com o trabalho desenvolvido por MITCHELL (1997), é um campo dentro da Inteligência Artificial (IA) que tem como objetivo desenvolver algoritmos capazes de resolver problemas complexos, identificando padrões e sendo capaz de realizar previsões ou tomar decisões baseado em conjuntos de dados (*datasets*). Desde que começou a ser formalizado como um campo de estudo, teve como meta a criação de sistemas que possam aprender automaticamente a partir de exemplos, sem a necessidade de programação explícita para cada tarefa, revolucionando a forma como problemas computacionais poderiam ser abordados (DOMINGOS, 2017).

Os algoritmos de *Machine Learning* são treinados com grandes conjuntos de dados para identificar padrões e relações, após esse treinamento, devem ser capazes de generalizar o conhecimento adquirido e, dessa forma, prever ou classificar novos dados (ZHOU, 2021). O *machine learning* será usado nesta pesquisa para treinar modelos de previsão da produção agrícola com base nas variáveis climáticas e nos dados históricos de produção. O foco será em como esses algoritmos podem identificar padrões e relações entre o clima e a produtividade agrícola, os algoritmos utilizados são apresentados nas seções *Random Forest*, *XGBoost* e *LSTM*.

A escolha dos algoritmos de aprendizado de máquina nesta pesquisa foi feita com base na necessidade de abordar diferentes aspectos do problema: identificação de variáveis climáticas mais relevantes e previsão da produtividade agrícola futura sob cenários climáticos distintos. A complexidade do contexto agrícola, que envolve múltiplas variáveis com comportamento não linear e sazonal, exige uma abordagem combinada com diferentes tipos de modelos. Os algoritmos Random Forest e XGBoost foram utilizados para identificar e ranquear a importância das variáveis climáticas na produção de soja e arroz. O Random Forest destaca-se por oferecer interpretabilidade por meio da análise de importância das variáveis. Já o XGBoost, baseado em boosting, é adequado para cenários com muitas variáveis e interações complexas. Mais detalhes do funcionamento destes algoritmos são apresentados nas próximas seções.

No entanto, esses modelos não são projetados para lidar com dependências temporais de longo prazo, característica essencial dos dados agrícolas utilizados nesta pesquisa. Para essa finalidade, foi adotada a arquitetura LSTM, que foi desenvolvida para resolver os problemas de desvanecimento e explosão do gradiente comuns em RNNs simples, sendo capaz de preservar informações ao longo do tempo, o que a torna ideal para a modelagem de séries temporais longas e com forte sazonalidade, como é o caso das safras agrícolas.

Modelos estatísticos clássicos como o ARIMA (BOX; JENKINS; REINSEL, 2008) também foram considerados, porém não foram adotados por apresentarem limitações importantes no contexto deste trabalho. O ARIMA pressupõe que as séries sejam estacionárias, o que pode comprometer a interpretação dos dados. Além disso, sua capacidade de modelar relacionamentos multivariados não lineares é restrita, tornando-se menos eficaz diante da complexidade das interações entre variáveis climáticas e produção agrícola.

Portanto, a adoção combinada dos algoritmos Random Forest, XGBoost e LSTM se mostrou essencial: os dois primeiros para a explicação do passado, indicando quais variáveis climáticas mais impactaram a produtividade e o último para a previsão futura, considerando os diferentes cenários climáticos projetados pelo modelo CMIP6. Essa abordagem integrada oferece interpretação e previsibilidade, características que nenhum dos modelos, isoladamente, conseguiria proporcionar de maneira satisfatória.

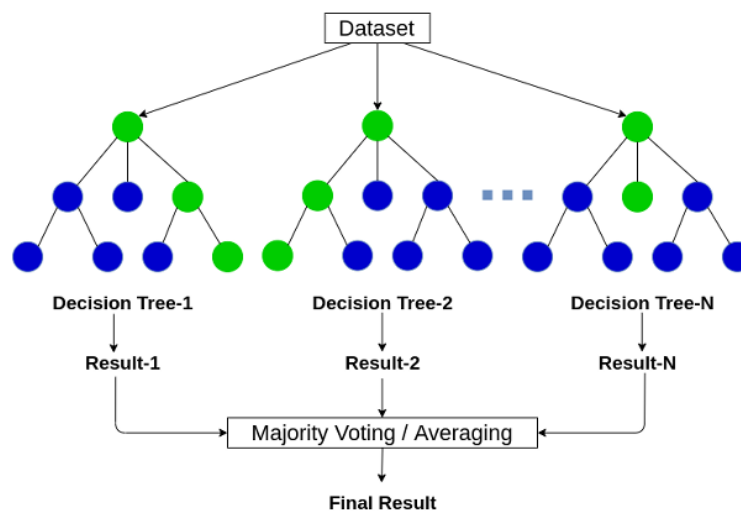
2.5.1 Random Forest

É um algoritmo de *machine learning* utilizado principalmente para tarefas de regressão e classificação, ele se encaixa na categoria dos algoritmos **supervisionados** que aprendem a partir de dados rotulados. A técnica consiste em gerar múltiplas árvores de decisão durante o treinamento, seu objetivo é combinar a saída destas árvores para obter um único resultado, a saída, por sua vez, pode ser a média (em casos de regressão) ou o voto majoritário (em casos de classificação) das árvores individuais (BREIMAN, 2001).

Com a técnica de *Bagging* proposta por BREIMAN (2001), cada árvore é treinada com diferentes subconjuntos de dados (amostras de treinamento) e variáveis selecionadas aleatoriamente, pertencentes ao conjunto de dados original. Para cada amostra de treinamento, uma árvore de decisão é construída.

Uma das principais vantagens do *Random Forest* é sua capacidade de reduzir o risco de *overfitting*, que ocorre quando um modelo se ajusta excessivamente aos padrões dos dados de treinamento, incluindo ruídos e detalhes irrelevantes, tornando-se incapaz de generalizar bem para novos dados. São utilizados hiperparâmetros para controlar o processo de treinamento como o número de árvores (afeta a precisão e o tempo de treinamento), a profundidade máxima da árvore (define a complexidade modelo), e o número mínimo de amostras (auxilia no controle do overfitting). Além do mais, o algoritmo fornece uma estimativa da importância de cada variável para o modelo, recurso muito útil em tarefas de seleção de atributos. A seguir, é apresentado na Figura 7 o esquema de funcionamento do *Random Forest*.

Figura 7 – Esquema do Random Forest.



Fonte: (TECHNOLOGIES, 2024)

2.5.2 XGBoost

O *XGBoost* (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina do tipo **supervisionado**, utilizado para tarefas de regressão e classificação. Ele se baseia na técnica de *boosting*, que constrói modelos sequenciais de árvores de decisão, onde cada nova árvore busca corrigir os erros cometidos pelas árvores anteriores (CHEN; GUESTRIN, 2016). Ao contrário do *Bagging*, o *boosting* dá mais peso às instâncias mal

classificadas, permitindo que o modelo aprenda com os erros e melhore progressivamente o desempenho.

Conhecido por sua alta eficiência e precisão, é otimizado para velocidade e desempenho computacional. O *XGBoost* incorpora regularização L1 e L2 para reduzir o risco de *overfitting*, também permite paralelização no treinamento, suporte a dados esparsos e tratamento automático de valores ausentes. Entre os hiperparâmetros deste algoritmo, estão a taxa de aprendizado (*learning rate*), o número de árvores, a profundidade máxima das árvores além do parâmetro de regularização. Esses recursos tornam o algoritmo uma escolha eficaz para problemas com grande volume de dados e múltiplas variáveis preditoras.

Outra característica do *XGBoost*, e motivo para sua utilização nesta pesquisa, é o fato de fornecer uma métrica de importância das variáveis utilizadas, semelhante ao *Random Forest*, permitindo avaliar o impacto relativo de cada atributo na predição final, sua aplicação se torna uma alternativa ao Random Forest para definição da importância de cada variável climática no período de safra tanto na série temporal quanto nas predições futuras. Devido à sua eficácia, o algoritmo tem sido amplamente adotado em competições de ciência de dados e aplicações reais em diferentes domínios.

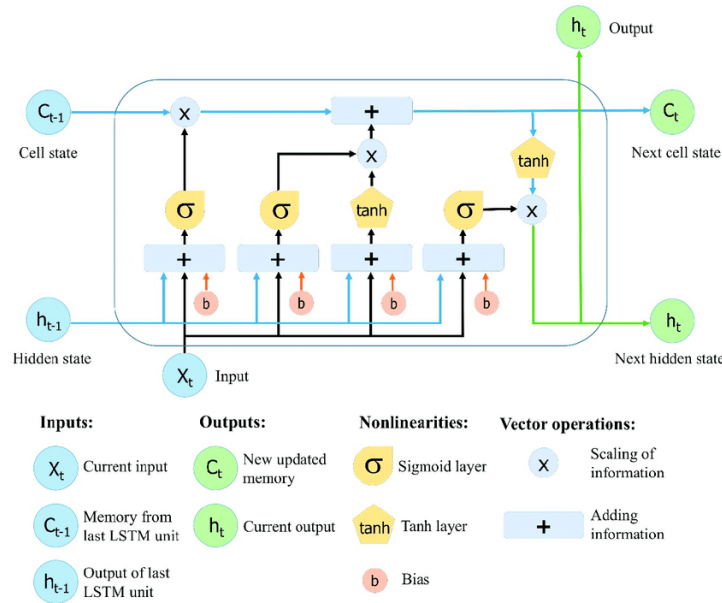
2.5.3 Long Short-Term Memory (LSTM)

As redes LSTM são uma variação das *Recurrent Neural Network*, e foi projetada para aprender dependências de longo prazo em sequências de dados (HOCHREITER; SCHMIDHUBER, 1997). As RNNs são uma classe de redes neurais artificiais projetadas para lidar com dados sequenciais no qual a ordem das entradas é relevante. Elas possuem conexões recorrentes e isso permite que informações de entradas anteriores influenciem no processamento da entrada atual, dessa forma, as RNNs se tornam adequadas para tarefas como modelagem de séries temporais.

As RNNs enfrentam limitações no aprendizado de dependências de longo prazo, causado por problemas como desaparecimento e explosão do gradiente (PHILIPP; SONG; CARBONELL, 2017) durante o treinamento. Isso dificulta a captura de padrões que dependem de eventos passados muito distantes na sequência.

Para superar essas dificuldades, as LSTMs apresentam uma arquitetura composta por células de memória, que contém mecanismos de controle conhecidos como portas (*input gate*, *forget gate* e *output gate*), que regulam o fluxo de informações ao longo do tempo, por meio de funções de ativação, geralmente funções sigmoide e tanh (LE et al., 2019). Essas portas permitem que a rede decida o que deve ser feito com as informações, armazenamento, atualização ou descarte, possibilitando a preservação de informações relevantes por períodos mais longos. A Figura 8 apresenta a arquitetura de uma célula no modelo LSTM.

Figura 8 – Arquitetura de uma célula no modelo LSTM.



Fonte: (YAN, 2016)

2.6 MÉTODOS DE AVALIAÇÃO

A utilização de métricas serve para avaliar a precisão de modelos preditivos. Entre as mais comuns, estão, Raiz do Erro Quadrático Médio (RMSE), Erro Percentual Absoluto Médio (MAPE) e o Coeficiente de Determinação (R^2), as quais foram utilizadas neste trabalho e serão apresentados a seguir.

2.6.1 RMSE

Root Mean Square Error, traduzido como Raiz do Erro Quadrático Médio (RMSE), tem sido usada como métrica estatística padrão na avaliação do desempenho de modelos em estudos meteorológicos, de qualidade do ar e pesquisas climáticas (CHAI; DRAXLER, 2014). Nela, é calculada a média dos erros ao quadrado entre os valores previstos e os valores reais, na sequência, aplica a raiz quadrada no resultado obtido. O RMSE caracteriza-se pela penalização com maior intensidade de grandes erros, se tornando útil quando grandes desvios devem ser evitados. Sua fórmula é a seguinte:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.1)$$

A variável y_i representa o valor observado, enquanto $\hat{y}_i$ é o valor previsto e n o número de observações. Uma das vantagens do RMSE é o fato do resultado estar expresso na mesma unidade de medida da variável utilizada na predição, ainda, segundo CHAI; DRAXLER (2014), é uma métrica robusta e amplamente utilizada em problemas de regressão e séries temporais, se encaixando na proposta desta pesquisa.

2.6.2 MAPE

Mean Absolute Percentage Error, traduzido para o português como Erro Percentual Absoluto Médio (MAPE) é capaz de expressar o erro médio das previsões em valores percentuais, o que o torna útil na interpretação do desempenho do modelo independente da escala e unidade de medida do conjunto de dados, uma abordagem oposta ao RMSE, como apresentado acima. Quanto menor o valor do MAPE, maior a acurácia do modelo. Essa métrica é comumente utilizada em estudos de previsão por sua facilidade de interpretação, especialmente em séries temporais e regressão (HYNDMAN; KOHLER, 2006), sua equação é dada por:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.2)$$

Sendo n o número total de observações, y_i o valor real observado e $\hat{y}_i$ o valor previsto. Como essa métrica mede o tamanho do erro, o valor absoluto (módulo) é obrigatório, possui facilidade de interpretação. Porém, o MAPE apresenta limitações quando os valores reais se aproximam ou são iguais a zero (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018), devido ao fato do valor real observado estar presente no denominador da equação. Desse modo, cidades que não possuem dados de produção agrícola nos cultivos estudados em determinados anos e conseqüentemente, apresentam valores iguais a zero, foram desconsideradas para utilização desta métrica.

2.6.3 R^2

O Coeficiente de Determinação (R^2) mede o quanto as previsões do modelo se ajustam aos valores reais observados, seu valor varia entre 0 e 1. Um coeficiente de determinação igual a 1 significa que o modelo é capaz de explicar 100% da variabilidade dos dados, isso significa dizer que todas as previsões se ajustam perfeitamente aos valores observados. Em contrapartida, um coeficiente igual a 0, significa que o modelo não conseguiu extrair nenhum padrão dos dados, e suas previsões serão equivalentes a média dos valores observados, conforme apresentado em MONTGOMERY; PECK; VINING (2012).

Sua fórmula é dada por:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.3)$$

Sendo y_i o valor real observado, $\hat{y}_i$ o valor previsto, n o número total de observações e $\bar{y}$ a média entre todos os valores reais observados.

2.7 MODELO CLIMÁTICO CMIP6

A *World Climate Research Programme* (WCRP), é a mais antiga e única iniciativa dedicada exclusivamente à coordenação de pesquisas climáticas internacionais, e que por sua vez tem desenvolvido conhecimentos fundamentais sobre a variabilidade do sistema climático, permitindo uma compreensão preditiva e chegando a conclusão de que a atividade humana é responsável pela maioria das mudanças climáticas observadas (World Climate Research Programme, 2024). Esse programa possui patrocínio da Organização Meteorológica Mundial (OMM), do Conselho Internacional para a Ciência (ICSU), e também da Comissão Oceanográfica Intergovernamental (COI). Desde sua fundação no ano de 1980, a WCRP tem sido fundamental para impulsionar o avanço da ciência climática, promovendo a oportunidade aos cientistas do clima de monitorar, simular e prever o clima global com alta precisão.

Como fruto desse programa, surgiu o modelo CMIP6, um conjunto padronizado de experimentos de simulação climática desenvolvido por centros de pesquisa de diferentes países, esse modelo fornece uma base comum de comparação de diferentes modelos climáticos, permitindo avaliações de incertezas e melhora nas previsões, segundo EYRING et al. (2016). Tem servido como fonte para a geração de relatórios do Painel Intergovernamental sobre Mudanças Climáticas (IPCC). O modelo CMIP6 apresenta diferentes projeções climáticas, conhecidas como cenários, que envolvem questões ambientais, sociais e econômicas como base para cada resultado, os cenários são explicados de forma detalhada na próxima seção.

2.7.1 Cenários SSP

Os cenários SSP (*Shared Socioeconomic Pathways*) são projeções que servem para analisar o impacto das mudanças climáticas em cenários de desenvolvimento socioeconômicos e emissões de gases de efeitos distintos. Foram desenvolvidos para fornecer uma base para os modelos climáticos de avaliação, são consideradas variáveis como crescimento populacional, desenvolvimento econômico e políticas climáticas, de acordo com

O'NEILL et al. (2017). Cada cenário reflete diferentes trajetórias de emissões, umas mais otimistas com políticas de redução de emissões, outras pessimistas com ausência das mesmas e alta emissão de gases poluentes. A seguir serão apresentados quatro cenários que são utilizados nesta pesquisa com suas características, utilizando como base o estudo de RIAHI et al. (2017). Em relação a nomenclatura dos cenários, o primeiro dígito representa o tipo de cenário socioeconômico, enquanto os últimos 2, representam o nível de forçamento radioativo previsto para o ano de 2100, isto é, quanto o planeta vai reter de energia extra (medida em *Watt*), em forma de radiação por metro quadrado na superfície da Terra, devido a gases de efeito estufa, sua unidade de medida é expressa em:

$$\text{W/m}^2 \quad (2.4)$$

Os dois últimos dígitos presentes na nomenclatura do cenário são expressos com uma vírgula decimal, como exemplo, o cenário SSP 126 descrito a seguir, representa o cenário socioeconômico 1 e um forçamento radioativo de 2,6 W/m².

- **SSP 126:** É caracterizado por um baixo nível de emissões de gases do efeito estufa e aumento limitado da temperatura global, promovidos pelo desenvolvimento sustentável, resultante de políticas climáticas eficazes;
- **SSP 245** Representa um cenário intermediário, possuindo emissões moderadas e políticas um pouco menos eficientes em comparação ao cenário SSP 126, nele, é considerada uma transição gradual para uma economia de baixo carbono;
- **SSP 370:** É um cenário de altas emissões, no qual o crescimento econômico avança sem destaque para políticas relacionadas às mudanças climáticas, alto crescimento populacional e uso intensivo de combustíveis fósseis;
- **SSP 585:** Este é o cenário mais pessimista, possui o maior nível de emissões de gases do efeito estufa e é acompanhado por falta de políticas de mitigação, a temperatura global aumenta consideravelmente.

3 TRABALHOS RELACIONADOS

Diversos estudos tem sido realizados com o objetivo de prever a produtividade agrícola com ênfase nos cultivos de soja e arroz, utilizando técnicas de *Machine Learning* e estatística. A seguir, serão apresentados alguns desses estudos com uma análise de suas abordagens e resultados obtidos.

O trabalho intitulado "*A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast*" (OLIVEIRA et al., 2021) propôs o desenvolvimento de um sistema escalável de aprendizado de máquina para previsão da produtividade agrícola, sua região de estudo é composta por 1500 municípios do Brasil e dos EUA. O sistema é baseado em uma rede neural recorrente (RNN), treinada com dados de precipitação, temperatura e propriedades do solo. A arquitetura do sistema é composta por uma API REST que integra dados meteorológicos e de propriedades do solo, os dados são provenientes de fontes como satélites e modelos climáticos sazonais, permitindo previsões com até sete meses de antecedência. Os resultados demonstraram que as previsões de produtividade de soja obtidas pelo sistema em alguns casos, foram superiores as previsões de modelos que utilizam dados de sensoriamento remoto, a proposta se assemelha a este trabalho pela utilização de dados como precipitação, temperatura e solo, além do fato de utilizar uma RNN para realizar as previsões.

O estudo "Análise estatística e modelos de *machine learning* na produção agrícola brasileira: tendências temporais e eficiência produtiva ao longo de quatro décadas (1980 - 2019)", de VASCONCELOS; SILVA (2021, se propôs a analisar a evolução da produção agrícola no Brasil utilizando dados da Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) entre os anos de 1980 a 2019. A metodologia combinou técnicas estatísticas tradicionais (análise de Pareto, teste de Shapiro-Wilk e teste t de uma amostra), com modelos de *machine learning*, (*Random Forest* e *Support Vector Machines* - SVM), aplicados para prever a produção agrícola e identificar padrões históricos. Os resultados mostraram um crescimento robusto na produção de culturas como cana-de-açúcar e oleaginosas, refletindo a expansão do agronegócio e o avanço tecnológico no setor. O rendimento médio (produção por hectare) também apresentou melhora, especialmente no século XXI. Apesar disso, os modelos de *machine learning* apresentaram limitações preditivas, com altos valores de erro quadrático médio (MSE) e coeficientes de determinação (R^2) negativos, sugerindo a necessidade de incorporar variáveis climáticas e abordagens mais sofisticadas em futuros estudos. O trabalho de VASCONCELOS; SILVA, reforça a necessidade de utilização de variados dados climáticos, como foi usado neste trabalho, para que o modelo seja capaz de compreender e explicar a variabilidade dos dados.

BATISTELLA (2023) desenvolveu o trabalho "Estimativa da Produtividade da Soja Utilizando Redes Neurais Artificiais" no qual aplicou técnicas de aprendizado de máquina

supervisionado para prever a produtividade da soja no Núcleo Regional de Pato Branco (PR) durante as safras de 2019/2020, 2020/2021 e 2021/2022. Sua base de dados foi composta por índices de vegetação extraídos do sensor MSI do satélite de observação da Terra pertencente ao programa *Copernicus*, Sentinel-2, além de dados como precipitação acumulada, altitude, tipo de solo, geologia e declividade. Os modelos testados incluíram RNAs com uma camada oculta, *Random Forest* e *Support Vector Machines* de regressão. A metodologia envolveu a separação dos dados em conjuntos de treinamento e teste, utilizando validação cruzada para avaliar a robustez dos modelos. O desempenho dos algoritmos foi comparado por meio de métricas como o coeficiente de determinação (R^2) e raiz do erro quadrático médio (RMSE). Os resultados indicaram que o Random Forest obteve o melhor desempenho geral, com valores de R^2 superiores a 0,80, seguido pelas RNAs e SVMs. Além da utilização do algoritmo *Random Forest*, BATISTELLA, também utiliza em sua abordagem duas métricas de avaliação usadas nesta pesquisa, coeficiente de determinação e raiz do erro quadrático médio.

Possuindo uma região de estudo mais reduzida, o trabalho de ARSEGO (2017), intitulado "Modelo estatístico de previsão de produtividade de soja e arroz para o Rio Grande do Sul", é o que mais se aproxima da abordagem utilizada no desenvolvimento da pesquisa aqui apresentada. Arsego, propôs um modelo de regressão linear múltipla para prever a produtividade de soja e arroz no estado, com base em dados históricos climáticos e de produtividade fornecidos pelo IBGE e INMET, abrangendo todo o estado do Rio Grande do Sul, as variáveis utilizadas no modelo incluíram precipitação acumulada, temperatura média, umidade relativa do ar e radiação solar correlacionados com registros históricos de produção de soja e arroz. Seu desempenho foi avaliado por meio de métricas como o coeficiente de determinação (R^2) e RMSE. Os resultados indicaram que a precipitação e a temperatura média durante o período reprodutivo foram os fatores mais influentes na produtividade. O modelo apresentou um bom desempenho na previsão de safras passadas, sugerindo sua aplicabilidade na projeção da produtividade futura, porém, diferentemente da abordagem utilizada neste trabalho, Arsego não emprega o uso de modelos climáticos futuros para correlacionar com a produtividade agrícola.

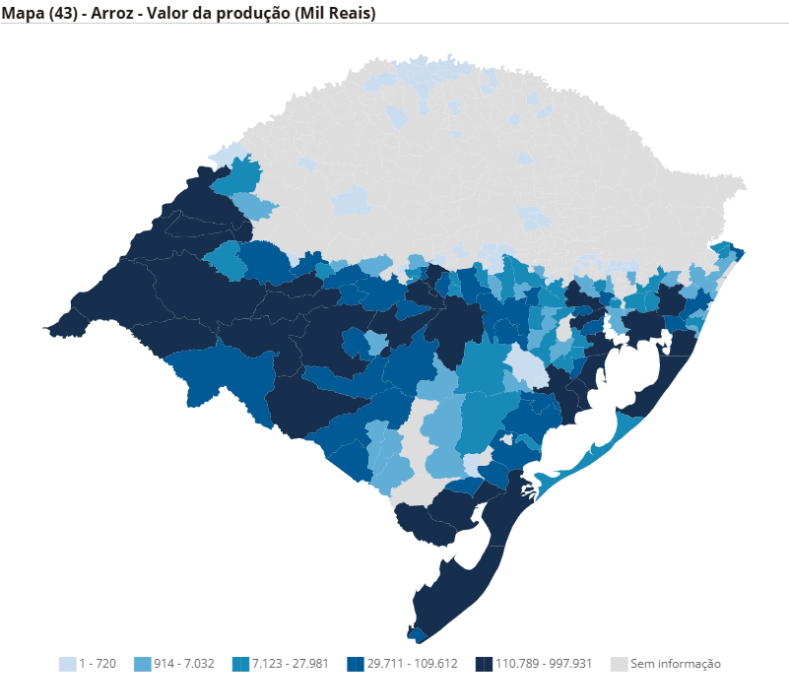
4 METODOLOGIA

Este capítulo descreve os procedimentos metodológicos adotados para o desenvolvimento desta pesquisa. Inicialmente, são apresentados os critérios utilizados para a seleção dos dados climáticos e agrícolas, seguidos pelas etapas de pré-processamento, como o tratamento de valores ausentes. Em seguida, aborda-se a associação entre os municípios da Região Geográfica Intermediária de Santa Maria e os dados meteorológicos. Por fim, são descritos os modelos de aprendizado de máquina utilizados para identificar padrões históricos e realizar previsões futuras, juntamente com as métricas aplicadas para avaliação de desempenho.

4.1 JUSTIFICATIVA DA REGIÃO DE ESTUDO

A Região Geográfica Imediata de Santa Maria (RS), apresentada na Figura 2, abrange um total de 25 municípios como mencionado no capítulo 2. Já a Região Geográfica Intermediária de Santa Maria, mostrada na Figura 1, possui um total de 40 municípios, mais abrangente, engloba as regiões imediatas de Santa Maria, Santiago, Cachoeira do Sul e São Gabriel - Caçapava do Sul. Estas duas últimas, é onde estão localizados municípios que se destacam como grandes produtores de arroz, conforme mostrado na Figura 9. Dessa forma, se torna a região mais adequada para o estudo, já que todos os municípios das regiões imediatas citadas se destacam na produção de soja, como podemos observar na Figura 10. Portanto, para uma análise envolvendo os dois cultivos, a Região Geográfica Intermediária de Santa Maria se torna a melhor opção.

Figura 9 – Produção de Arroz no RS em 2023.



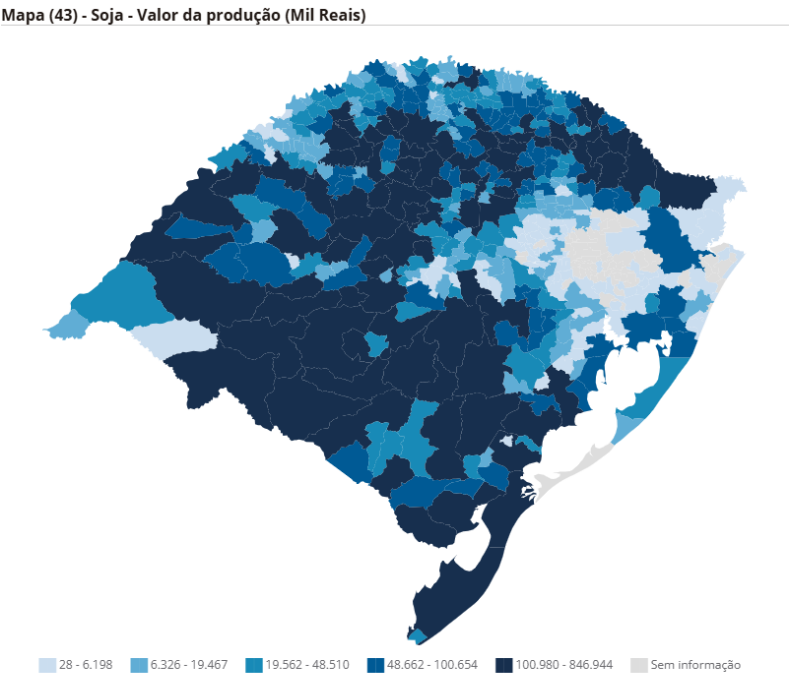
Fontes

[PAM](#): Valor da produção, Quantidade produzida, Área colhida, Rendimento médio, Maior produtor

[Censo Agropecuário](#): Estabelecimentos, Número de pés

Fonte: (Instituto Brasileiro de Geografia e Estatística (IBGE), 2025a)

Figura 10 – Produção de Soja no RS em 2023.



Fontes

[PAM](#): Valor da produção, Quantidade produzida, Área colhida, Rendimento médio, Maior produtor

[Censo Agropecuário](#): Estabelecimentos, Número de pés

Fonte: (Instituto Brasileiro de Geografia e Estatística (IBGE), 2025b)

4.2 SELEÇÃO E PRÉ-PROCESSAMENTO DE DADOS

Nesta seção serão apresentados os processos e as justificativas relacionadas a seleção e pré-processamento dos dados utilizados para a realização dessa pesquisa. Os *datasets* estão divididos em 4 áreas: municípios, produção agrícola, dados climáticos históricos e dados climáticos futuros. Após a seleção e obtenção desses *datasets*, os próximos passos dizem respeito ao pré-processamento, sendo apresentados nas subseções Associação de Estações Meteorológicas aos Municípios e Tratamento de Dados Ausentes e Agregações.

4.2.1 Dados dos Municípios

Inicialmente, foram filtrados os 40 municípios pertencentes a região de estudo, de acordo com dados do IBGE e seus nomes foram adicionados em uma planilha, posteriormente foram obtidas as coordenadas associadas a cada município de acordo com o IBGE. Para isso, foi utilizado um repositório¹ disponível no Github que utiliza a API do IBGE², a qual disponibiliza diversas informações sobre as localidades do Brasil. Para isso, foi criado um algoritmo em Python que lê o nome dos municípios, contidos na planilha e filtra os dados das variáveis de latitude e longitude que estão presentes no arquivo CSV contido no repositório, utilizando o nome da cidade e o código da unidade federativa, na sequência, o algoritmo adiciona na planilha os dados das respectivas coordenadas geográficas de cada município, finalizando a etapa de obtenção de dados.

4.2.2 Dados de Produção Agrícola

O IBGE realiza uma pesquisa anual sobre a produção agrícola, o estudo se chama Produção Agrícola Municipal e contém diversos dados relacionados à agricultura nas esferas municipal, estadual, regional e federal. Os resultados desde que a pesquisa começou a ser realizada, no ano de 1973, estão situados no sistema SIDRA, cuja sigla significa Sistema IBGE de Recuperação Automática, neste sistema é possível selecionar os tipos de variáveis relacionadas a produção agrícola, cultivos, período de tempo e nível territorial. Para esta pesquisa foram selecionadas as variáveis da produção agrícola: quantidade produzida (toneladas) e rendimento médio (kg/m^2), o nível territorial selecionado foi a Região Geográfica Intermediária de Santa Maria (RS), dessa forma, foram filtrados os 40 municípios pertencentes a esta região, os cultivos selecionados foram a soja (em grãos) e o

¹<https://github.com/kelvins/municipios-brasileiros/>

²<https://servicodados.ibge.gov.br/api/docs/localidades>

arroz (em casca), o processo foi realizado duas vezes, um para cada cultivo em separado, gerando assim dois *datasets* relacionados a produção agrícola, o período de tempo definido foi dos anos de 1994 a 2023, o que significa, o dados de 30 safras de cada cultivo na região de estudo.

4.2.3 Dados Climáticos Históricos (INMET e NASA POWER)

A escolha das variáveis climáticas utilizadas neste estudo: **temperatura, precipitação, radiação solar e umidade do solo** - fundamenta-se em estudos científicos que relatam sua influência direta sobre o desenvolvimento e a produtividade de culturas como arroz e soja. A temperatura é um dos principais reguladores dos estágios fenológicos das plantas, afetando processos como germinação, florescimento e enchimento de grãos (FEHR; CAVINESS, 1977; COUNCE; L.; FEHR, 2000). A precipitação, define a disponibilidade hídrica para as culturas, sendo essencial que sua distribuição ocorra em momentos críticos do ciclo para evitar déficits que comprometam a produtividade (CUNHA; WREGGE; MALUF, 2001; SENTELHAS; BATTISTI; CÂMARA, 2015). A radiação solar representa a principal fonte de energia para a fotossíntese, e sua disponibilidade ao longo do ciclo impacta diretamente a produção de biomassa e o rendimento final (MONTEITH, 1977; BERGAMASCHI et al., 2004). Já a umidade do solo funciona como um indicador do armazenamento hídrico disponível para as plantas, sendo muito importante em regiões com alta variabilidade pluviométrica, como é o caso do Rio Grande do Sul (OLIVEIRA; CUNHA; STRECK, 2015; PEREIRA; ANGELOCCI; SENTELHAS, 2002). Dessa forma, a seleção dessas variáveis tem por objetivo, capturar os principais fatores climáticos que afetam o desempenho agrícola na região de estudo.

O INMET possui estações meteorológicas espalhadas por todo o país, elas estão divididas entre estações convencionais e automáticas. Mais antigas, as convencionais utilizam equipamentos mecânicos e requerem leitura manual dos dados por meio de um observador, possuem uma longa série histórica do clima. Já as estações meteorológicas automáticas, utilizam sensores eletrônicos que registram e transmitem dados de forma automática, seus dados são mais precisos, porém são mais recentes. Dessa forma, foram selecionadas as estações meteorológicas convencionais por apresentarem dados mais antigos e necessários para se alcançar os objetivos dessa pesquisa. Os dados foram obtidos pelo sistema BDMEP³ - Banco de Dados Meteorológicos par Ensino e Pesquisa do INMET, no qual após fornecer um endereço de *email* pelo qual será enviado o arquivo contendo o *dataset* após o processamento por parte do sistema, foram utilizadas as definições e filtros apresentados no Quadro 1. Os dados meteorológicos do INMET não contém todas as variáveis necessárias para este estudo, portanto, foram selecionados apenas as variá-

³<https://bdmep.inmet.gov.br/>

veis precipitação total e temperatura média compensada. Como resultado, obteve-se um arquivo CSV contendo os dados selecionados para cada um das estações convencionais do estado do Rio Grande do Sul, totalizando 25 estações, além dos dados meteorológicos, estão contidas informações como nome da cidade na qual a estação está localizada, coordenadas geográficas, ano de início de atividades, situação (ativa ou desativada, caso esteja desativada, apresenta a data do fim da atividade da estação). Vale ressaltar que muitas delas tornaram-se inativas ao longo do tempo, além disso, possuem ausência de dados ao longo de determinados períodos.

Quadro 1 – Variáveis selecionadas no sistema BDMEP - INMET.

Tipo de Pontuação:	Ponto
Tipo de Dado:	Dados Mensais
Tipo de Estação:	Convencionais
Abrangência:	Região
Data de Início:	01/01/1993
Data de Fim:	31/12/2023
Regiões:	Sul
Variáveis:	Precipitação Total, Temperatura Média Compensada
Estações:	Todas as estações do estado do RS

Fonte: Do autor.

Para completar as variáveis climáticas necessárias, foi necessário encontrar os dados de radiação solar e umidade do solo em outra fonte. Foi utilizada a plataforma NASA POWER, por meio da ferramenta *Data Access Viewer*. Foram selecionadas as coordenadas geográficas que abrangem longitudinalmente o estado de leste a oeste e latitudinalmente de norte a sul, respeitando seus limites geográficos. A comunidade científica escolhida foi "*Agroclimatology*", com nível temporal mensal e anual, abrangendo o período de 1993 a 2023. Os parâmetros selecionados foram equivalentes aos utilizados nos modelos climáticos do CMIP6, sendo eles: "*ALLSKY_SFC_SW_DWN*", que corresponde à radiação solar de onda curta na superfície (equivalente ao parâmetro **rsds** do CMIP6), e "*SOILM*", que representa o conteúdo volumétrico de umidade do solo superficial (equivalente ao parâmetro **mrsos**). Os dados foram exportados em formato NetCDF para posterior pré-processamento e integração ao modelo de análise, os parâmetros **rsds** e **mrsos** são apresentados na próxima seção.

4.2.4 Dados Climáticos Futuros (CMIP6)

Como conjunto de dados climáticos futuros, são utilizados 4 cenários do modelo CMIP6: SSP 126, SSP 245, SSP 370 e SSP 585 - abordados no capítulo 2, dessa forma,

foi utilizado o sistema da *Earth System Grid Federation*⁴ (ESGF), uma parceria de centros de modelagem climática dedicados a apoiar a pesquisa climática (NASA Center for Climate Simulation, 2024).

Para selecionar os dados climáticos futuros, foi necessário compreender o significado de cada variável presente no sistema, há um repostório no github chamado WCRP-CMIP⁵ criado pelo *Program for Climate Model Diagnosis and Intercomparison* (PCMDI) que explica o significado das variáveis e a composição do nome de cada modelo, portanto, o repositório serviu de base para compreensão e definições de filtros a serem aplicados.

O modelo selecionado foi o **EC-Earth3**, desenvolvido por um consórcio europeu é uma evolução do modelo climático do *European Centre for Medium-Range Weather Forecasts* (ECMWF), adaptado para simulações de longo prazo do sistema terrestre. Para cada variável selecionada, foi utilizada a simulação identificada por *variant label*: **r1i1p1f1**. Essa notação indica a versão específica da execução do modelo, no qual:

- **r1**: refere-se à primeira realização (*realization*), ou seja, a primeira execução do modelo com pequenas perturbações nas condições iniciais;
- **i1**: indica o primeiro conjunto de condições iniciais (*initialization*);
- **p1**: representa a primeira configuração de parâmetros físicos (*physics*) utilizados no modelo;
- **f1**: representa o primeiro conjunto de forçantes externas (*forcing*), como concentrações de gases de efeito estufa ou aerossóis.

A malha espacial dos dados é definida por *grid label*: **gr** - que corresponde a uma grade regular interpolada a partir da grade original do modelo, sendo esta a mais comum e amplamente utilizada para análises espaciais. As variáveis **pr** (precipitação), **rsds** (radiação solar na superfície), e **tas** (temperatura do ar próxima à superfície) estão contidas na tabela **Amon**, que agrupa variáveis atmosféricas mensais. Já a variável **mrsos**, que representa a umidade do solo na camada superficial, está incluída na tabela **Lmon**, que contém variáveis mensais da superfície terrestre.

Para obtenção dos dados, observou-se a necessidade da aplicação dos filtros na exata ordem dos parâmetros mostrados na coluna da esquerda no Quadro 2, variações na ordem de aplicação de filtros, resultaram na falta de resultados encontrados. Entre parênteses, as diferentes variáveis que foram selecionadas, sendo uma por vez, repetindo o processo até a obtenção de todos os *datasets* necessários. Por fim, foram baixados os arquivos NetCDF (*Network Common Data Form*) após aplicação dos filtros, são divididos anualmente, dessa forma foi realizado o *download* de 10 arquivos.

⁴<https://aims2.llnl.gov/search/cmip6/>

⁵https://github.com/WCRP-CMIP/CMIP6_CVs

Quadro 2 – Ordem de aplicação dos filtros no sistema do ESGF.

Variant label:	r1i1p1f1
Grid label:	gr
Table ID:	(Amon, Lmon)
Variable ID:	(pr, rsds, tas, mrsos)
Source ID:	EC-Earth3
Experiment ID:	(ssp126, ssp245, ssp370, ssp585)

Fonte: Do autor.

4.2.5 Tratamento de Dados Ausentes

Entre todos os dados coletados, a ausência de dados ocorreu apenas nos *datasets* de histórico climático das estações meteorológicas. Observou-se que haviam 4 casos distintos: ausência de dados em um período anterior ao início das atividades de determinada estação, ausência de dados em um período após o fim das atividades, ausência de dados em meses isolados, ausência de dados em meses sequenciais.

Nos dois primeiros casos, não foi realizada nenhuma ação de tratamento. Quando a ausência de dados ocorre em meses isolados, ou seja, quando o mês anterior e o posterior possuem valores disponíveis, foi aplicada a média entre esses dois meses, atribuindo o valor resultante ao mês faltante, tanto para a precipitação mensal quanto para a temperatura média mensal. Já nos casos em que há ausência de dados em meses consecutivos, o valor de cada mês ausente foi substituído pela média de todos os mesmos meses (janeiros, fevereiro, etc.) ao longo dos anos disponíveis naquela estação meteorológica. Para exemplificar, se os dados de janeiro e fevereiro de um determinado ano estão ausentes, o valor atribuído a janeiro será a média de todos os janeiros disponíveis, e o mesmo ocorrerá para fevereiro. Esse tratamento foi aplicado às duas variáveis obtidas pelas estações e implementado por um algoritmo desenvolvido pelo autor na linguagem Python.

4.2.6 Associação de Dados Climáticos e Agrícolas aos Municípios

Com os dados climáticos pré-processados, o próximo passo foi sua associação com cada município da região de estudo, no que diz respeito aos dados históricos, para associação dos dados do INMET aos municípios, foi utilizado o cálculo da distância de Haversine entre as coordenadas dos municípios e das estações meteorológicas. Cada município foi associado à estação meteorológica mais próxima, caso a mesma não apresentasse dados em um determinado período, foi utilizada a segunda mais próxima, e assim por diante até preencher todas as informações para determinado ano e município.

No caso dos dados climáticos históricos obtidos pelo NASA POWER, foi utilizado o método *nearest* da biblioteca Xarray para associar cada município ao ponto mais próximo

da grade do modelo. Em seguida, foi novamente calculada a distância de Haversine entre as coordenadas do município e a localização climática associada, tendo como único objetivo, a verificação da proximidade dos pontos da grade dos dados com cada município, a média de distância foi de 23,5Km, sendo a maior distância equivalente a 36,36Km, relacionada a cidade de São Vicente do Sul, enquanto a menor distância foi de 11,24Km em relação a cidade de Mata. Também foi utilizada a API do Geoapify para realizar a conferência geográfica dos pontos de grade climáticos. Para cada ponto selecionado nos dados climáticos históricos, foi extraído, o nome da localidade mais próxima, incluindo o estado e o país. Essa verificação teve como finalidade garantir que os dados climáticos utilizados fossem coerentes com a localização dos municípios da área de estudo.

Após, foram selecionados os meses relevantes para o ciclo de produção da soja e do arroz (outubro a março), o IBGE apresenta os dados das safras anualmente, considerando como ano da safra o da colheita, portanto, neste trabalho foi utilizada a mesma abordagem.

Os dados históricos foram agrupados anualmente para cada município, o ano, como mencionado anteriormente, sendo o da colheita, dessa forma foi gerado dois *dataframes* relacionados a soja, contendo: cidade, ano, dados climáticos históricos (precipitação, radiação solar, temperatura e umidade do solo) de outubro a março, e dados da produção agrícola da soja (quantidade produzida e rendimento médio) um deles contendo os dados com 30 anos (1994 a 2023) e outro com 20 anos (2004 a 2023). Os *dataframes* relacionados ao arroz possuem estrutura semelhante, com a diferença relacionada aos dados de produção que neste caso representam o cultivo de arroz, também com séries temporais de 20 e 30 anos. Sendo assim, cada *dataframe* com série temporal de 30 anos possui 1200 linhas, enquanto aqueles com série temporal de 20 anos possuem 800 linhas, cada uma representando um município e um ano de safra, além dos dados climáticos e agrícolas. Dessa forma na série temporal de 20 anos, para cada município, há 20 correspondências, enquanto na série de 30 anos, 30 correspondências, portanto, o número total de linhas é o produto entre o número total de cidades da região geográfica intermediária de Santa Maria/RS (40) e o número de correspondências de cada município no conjunto de dados (20 ou 30).

Além dos dados históricos, foram preparados quatro *Dataframes* para os cenários climáticos futuros, utilizando as projeções dos modelos climáticos do CMIP6 (SSP 126, SSP 245, SSP 370 e SSP 585). Esses dados foram organizados mantendo as mesmas variáveis mensais (precipitação, temperatura, radiação solar e umidade do solo) e associando aos 40 municípios da Região Geográfica Intermediária de Santa Maria/RS, a Tabela 1 apresenta as unidades de medida das variáveis. Para cada cenário, foi criado um *Dataframe*, com entradas para cada município e ano (2025-2034). Esses quatro *Dataframes* não incluem dados de produção agrícola, apenas variáveis climáticas, que serão utilizadas em conjunto com os modelos preditivos para realizar as previsões da produção agrícola.

Foram utilizadas ferramentas da linguagem de programação Python, como as bibliotecas Pandas, NumPy e Xarray. A biblioteca Pandas foi empregada para organizar os dados tabulares e realizar operações de junção e agregação. A Xarray foi utilizada para leitura e manipulação dos arquivos no formato NetCDF provenientes dos modelos climáticos CMIP6.

Tabela 1 – Unidades de medida das variáveis climáticas utilizadas

Variável	Unidade de Medida
Precipitação	Milímetros (mm). Representa a profundidade da lâmina de água precipitada sobre uma área durante um período de tempo.
Temperatura	Graus Celsius (°C). Indica a média da temperatura do ar ao longo do mês na superfície.
Radiação solar	Watts por metro quadrado (W/m ²). Refere-se à densidade de potência da radiação solar de onda curta que atinge a superfície terrestre, considerando céu claro e encoberto.
Umidade do solo	Quilogramas por metro quadrado (kg/m ²). Quantidade de água presente na camada superficial do solo, expressa como massa de água por área (equivalente a milímetros de água).

Fonte: Do autor.

4.3 MINERAÇÃO DE DADOS

Esta seção descreve a etapa de aplicação dos algoritmos de aprendizado de máquina para análise e previsão da produção agrícola, inserida no processo de Mineração de Dados dentro do processo KDD. Inicialmente, foram definidos os critérios de preparação dos dados, incluindo normalização, codificação categórica e divisão temporal entre conjuntos de treino e teste. Em seguida, foram desenvolvidos e avaliados modelos utilizando os algoritmos *Random Forest*, *XGBoost* e *LSTM*, com foco tanto na identificação da importância relativa das variáveis climáticas quanto na geração de previsões para diferentes culturas e cenários futuros. As análises consideraram séries temporais de 20 e 30 anos, referentes as safras históricas de arroz e soja nos municípios da Região Geográfica Intermediária de Santa Maria/RS. Esta abordagem possibilitou investigar, o papel das condições climáticas mensais sobre o rendimento médio e a quantidade produzida, contribuindo para a construção de modelos preditivos com boa capacidade de generalização.

4.3.1 Aplicação de Random Forest e XGBoost

Foram aplicados os algoritmos *Random Forest* e *XGBoost* com objetivo de identificar a importância de cada variável climática considerando todo o período de safra de cada cultivo, tanto para os dados históricos quanto para as previsões futuras. No primeiro momento, os algoritmos foram utilizados com datasets das últimas 20 e 30 safras de arroz e soja. Para cada um, foram desenvolvidos três modelos distintos, tendo como variáveis alvo o rendimento médio (Kg/ha) e a quantidade produzida (toneladas) em cada série temporal. Os dados climáticos e o ano da safra foram normalizados via *MinMaxScaler* para evitar influência desproporcional de variáveis com magnitudes diferentes, enquanto a variável categórica cidade foi representada por codificação *one-hot*. As variáveis preditoras incluíram dados climáticos mensais (precipitação, temperatura, radiação solar e umidade do solo), a safra normalizada e a codificação das cidades. A avaliação foi realizada utilizando as métricas RMSE, MAPE e coeficiente de determinação (R^2).

Antes do treinamento, os dados foram filtrados para remover registros com rendimento ou produção iguais a zero, além de remover cidades com menos de 10 safras. A separação de treino e teste foi feita respeitando a ordem temporal e estratificada por cidade: para cada cidade, os primeiros 80% dos anos foram usados para treino e os 20% finais para teste, garantindo que os modelos fossem avaliados para previsões futuras em municípios conhecidos.

Os três modelos de *Random Forest*, serão referenciados pelas suas respectivas siglas e foram configurados da seguinte forma: o primeiro (**RF1**) utilizou 300 árvores ($n_estimators=300$), profundidade máxima de 10 ($max_depth=10$), divisão dos nós com mínimo de duas amostras ($min_samples_split=2$), folhas com pelo menos uma amostra ($min_samples_leaf=1$), seleção de variáveis pela raiz quadrada do total ($max_features='sqrt'$) e amostragem com reposição ($bootstrap=True$). O segundo modelo (**RF2**) incorporou clusterização via *K-Means*, agrupando municípios em quatro clusters baseados no histórico de rendimento médio agrícola de todo o dataset, esses clusters foram codificados com one-hot e adicionados às variáveis preditoras, mantendo os demais hiperparâmetros do RF1. Já o terceiro modelo (**RF3**) aplicou a clusterização apenas aos dados históricos iniciais para evitar vazamento de informações futuras, utilizando rendimentos normalizados e incluindo a nova variável categórica da mesma forma que no modelo anterior, mantendo também os mesmos hiperparâmetros do RF1. Em todos os casos, a divisão temporal treino/teste respeitou a sequência cronológica para cada município.

Os modelos de *XGBoost* seguiram configuração semelhante: o primeiro (**XGB1**) foi configurado com 300 árvores ($n_estimators=300$), taxa de aprendizado 0,03, profundidade máxima 4 ($max_depth=4$), amostragem parcial de observações ($subsample=0.8$), penalizações L1 ($reg_alpha=0.5$) e L2 ($reg_lambda=0.8$), além de no mínimo de cinco instâncias por folha ($min_child_weight=5$). No segundo modelo (**XGB2**), os municípios foram agrupados em quatro clusters via *K-Means*, considerando o histórico de rendimento ou

produção de todo o dataset, com codificação one-hot incorporada às variáveis climáticas normalizadas, à safra e à cidade; a variável alvo foi transformada pelo logaritmo natural, mantendo os mesmos hiperparâmetros. O terceiro modelo (**XGB3**) aprimorou a clusterização ao aplicá-la somente aos dados históricos iniciais para evitar vazamento, usando rendimentos normalizados e integrando essa variável categórica via one-hot, junto com as demais variáveis normalizadas. A divisão treino/teste seguiu o mesmo critério temporal adotado nos modelos *Random Forest*.

Por fim, realizou-se uma análise da importância das variáveis com base nas divisões das árvores, avaliada sob duas perspectivas: soma da importância por tipo de variável climática e proporção normalizada das importâncias mensais de cada variável climática. Essa análise permitiu compreender o peso relativo dos fatores climáticos ao longo do ciclo produtivo de arroz e soja. A Tabela 2 apresenta de maneira sintetizada os modelos apresentados nesta seção.

Tabela 2 – Comparação entre os modelos Random Forest e XGBoost

Algoritmo	Modelo	Clusterização	Transformações adicionais	Configuração dos hiperparâmetros
Random Forest	RF1	Não	Nenhuma	<i>n_estimators=300, max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', bootstrap=True</i>
	RF2	K-Means (dados completos)	Codificação one-hot dos clusters	Mesmos hiperparâmetros do RF1
	RF3	K-Means (dados iniciais)	Codificação one-hot dos clusters	Mesmos hiperparâmetros do RF1
XGBoost	XGB1	Não	Nenhuma	<i>n_estimators=300, learning_rate=0.03, max_depth=4, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.5, reg_lambda=0.8, min_child_weight=5</i>
	XGB2	K-Means (dados completos)	One-hot dos clusters + <i>log</i> da variável alvo	Mesmos hiperparâmetros do XGB1
	XGB3	K-Means (dados iniciais)	Codificação one-hot dos clusters	Mesmos hiperparâmetros do XGB1

Fonte: Do autor.

4.3.2 Aplicação de LSTM

Como já descrito no capítulo 2, o LSTM foi escolhido para realizar a tarefa das previsões das safras futuras, devido a sua capacidade de capturar relações temporais em séries de dados e por ser especialmente adequado para contextos em que há dependência entre valores passados e futuros, no caso deste trabalho, essa dependência ocorre em relação ao clima nos meses de safra, além de um padrão de produtividade de cada município ao

longo das últimas décadas. Desse modo, foram desenvolvidos 3 modelos LSTM distintos, projetados para compreender essas relações e realizar predições das variáveis alvo rendimento médio e quantidade produzida dos cultivos de arroz e soja nas cidades da região de estudo, com séries temporais de 20 e 30 anos. Assim como na seção anterior, cada modelo será referenciado por sua respectiva sigla, o primeiro modelo apresentado, será representado pela sigla **LSTM1**, o segundo: **LSTM2**, terceiro: **LSTM3**.

Os três modelos propostos foram desenvolvidos com redes neurais bidirecionais, em todos os casos, os dados climáticos foram organizados em uma matriz com 6 meses (outubro a março) e 4 variáveis mensais (precipitação, temperatura, radiação solar e umidade do solo) resultando em uma entrada sequencial com dimensão (6,4). Esses dados foram normalizados com *MinMaxScaler*, tomando como base apenas o conjunto de treino para evitar vazamento de dados. Além da entrada sequencial, os modelos também incorporaram o ano da safra, representado como uma variável contínua e normalizada, o identificador do município foi processado por uma camada de *embedding* ao invés de *one-hot encoding*, tendo como objetivo a captura de similaridades latentes entre os municípios, para manter a dimensionalidade reduzida, a codificação via embeddings possui dimensão igual a 4. As variáveis alvo foram transformadas para escala logarítmica, para reduzir a assimetria e estabilizar a variância dos dados.

O treinamento de todos os modelos utilizou o otimizador *Adam* e a função de perda MSE (erro quadrático médio), com validação temporal baseada nos últimos 15% dos dados de treino, respeitando a ordem cronológica, novamente, para evitar vazamento de dados. Também foi aplicada a técnica de *EarlyStopping* com "*patience=30*" para interromper o treinamento quando não houvesse melhora na validação. A divisão dos dados seguiu a ordem temporal, para cada município que produziu em pelo menos 10 safras, as 80% primeiras safras foram usadas para treino enquanto as 20% finais foram utilizadas para teste. Os dados de treino foram agrupados em um único conjunto, visando aumentar a variabilidade e robustez do conjunto de aprendizado.

A arquitetura do modelo LSTM1 foi implementada utilizando o *Keras*, a entrada sequencial foi processada por duas camadas LSTM bidirecionais com 64 neurônios cada, a primeira com "*return_sequences=True*" e a segunda com "*return_sequences=False*". Entre essas camadas foi aplicado *dropout* de 30%, ou seja, a cada iteração, esse percentual de neurônios é desligado aleatoriamente, ajudando a reduzir o *overfitting*. Após, a saída sequencial é concatenada com um vetor representando o município (via *Embedding*, seguido de *Flatten*) e com o ano da safra normalizado. Esse vetor combinado passa por duas camadas densas com 64 e 32 neurônios, ambas com ativação ReLU, sendo a primeira regularizada com L2 igual a 0.0001 para penalizar pesos excessivamente grandes, contribuindo também para controle de *overfitting*. A camada de saída é uma *Dense* linear com um único neurônio, que gera a previsão final. O modelo foi treinado com *batch size* de 32 e até 250 épocas.

O segundo modelo, LSTM2, manteve a base conceitual do primeiro modelo, com duas camadas LSTM bidirecionais, mas reduz o número de neurônios da segunda camada para 32, criando uma arquitetura um pouco mais compacta para extração de características sequenciais. O *dropout* de 30% e a *Batch Normalization* continuam entre as camadas para garantir regularização e estabilidade, mas a principal diferença está na simplificação do LSTM2, buscando reduzir a complexidade do modelo e melhorar a generalização. O *embedding* para a cidade e a normalização do ano da safra permanecem iguais, assim como a combinação dessas entradas antes das camadas densas. As camadas densas com 64 e 32 neurônios, ativação ReLU e regularização L2 também se mantêm, assim como a camada final linear para saída. O treinamento segue os mesmos parâmetros, com *batch size* 32, até 250 épocas e *Early Stopping*.

No modelo LSTM3, a arquitetura é muito semelhante ao modelo LSTM2, utilizando duas camadas bidirecionais (com 64 e 32 unidades, respectivamente) seguidas por *dropout* de 30%, e concatena as saídas das LSTM com as informações da cidade e do ano antes de alimentar as camadas densas. No entanto, o modelo LSTM3 se diferencia por adotar uma abordagem mais robusta de regularização, incorporando camadas de *Batch Normalization* após as LSTM e camada densa, além de utilizar uma penalização L2 mais intensa, equivalente a 0.001, para reduzir o risco de *overfitting*. A seguir a Tabela 3 apresenta uma comparação entre os modelos LSTM utilizados.

Tabela 3 – Comparação entre os modelos LSTM1, LSTM2 e LSTM3

Característica	LSTM1	LSTM2	LSTM3
Camadas LSTM bidirecionais	2 (64, 64)	2 (64, 32)	2 (64, 32)
Uso de Dropout	Sim (30%)	Sim (30%)	Sim (30%)
Uso de BatchNormalization	Não	Sim (entre camadas LSTM)	Sim (entre LSTM e após Dense)
Embedding para cidade	Sim	Sim	Sim
Entrada do ano (normalizado)	Sim	Sim	Sim
Camadas densas	64 (ReLU, L2=0.0001), 32 (ReLU)	64 (ReLU, L2=0.0001), 32 (ReLU)	64 (ReLU, L2=0.001), 32 (ReLU)
Camada de saída	Dense linear (1 unidade)	Dense linear (1 unidade)	Dense linear (1 unidade)
Batch size	32	32	32
Épocas máximas	250	250	250
EarlyStopping	Sim	Sim	Sim
Regularização L2	0.0001	0.0001	0.001

Após a etapa de treinamento, foi realizada a geração das previsões para cada cenário climático e avaliação de desempenho dos modelos. Inicialmente, o modelo gerou as previsões para os conjuntos de treino e teste, utilizando as entradas normalizadas de clima, ano e cidade. Como as variáveis alvo foram previamente transformados com logaritmo e escalonamento, as previsões foram revertidas para a escala original por meio da inversão do escalonamento (*MinMaxScaler.inverse_transform*) e da inversão da transformação logarítmica (*np.exp*). Com os valores reais e predições devidamente ajustados, foram adicionadas colunas com valores das variáveis alvo ao final dos datasets de cada cenário climático SSP, dessa forma, esses conjuntos de dados, passaram a conter cidade, safra, dados climáticos e dados de rendimento médio e quantidade produzida de soja e arroz para cada município da região para as próximas 10 safras. Aos municípios que foram excluídos do treinamento por terem produzido em menos de 10 em cada série temporal, foi atribuído o valor zero para as variáveis alvo. Após, foram calculadas as métricas de avaliação RMSE, R^2 e MAPE, tanto para os dados de treino quanto para os de teste, com o objetivo de quantificar a precisão do modelo. Também foram gerados gráficos de dispersão para comparar os valores reais e preditos no conjunto de teste para complementar a análise numérica das métricas de avaliação de desempenho.

Ao final do processo, foram aplicados os algoritmos *Random Forest* e *XGBoost*

com os mesmos modelos descritos na seção anterior, com os novos conjuntos de dados de cada cenário SSP contendo as previsões, tendo por objetivo compreender a relação e a importância de cada variável climática na previsão para as próximas 10 safras de arroz e soja na Região Geográfica Intermediária de Santa Maria/RS.

4.4 INTERPRETAÇÃO DOS RESULTADOS

Como parte da metodologia deste trabalho e etapa da fase de interpretação do processo KDD, foi desenvolvido um sistema interativo para visualização e análise das projeções agrícolas sob diferentes cenários climáticos, utilizando a biblioteca *Streamlit* da linguagem de programação *Python*. A aplicação foi projetada para possibilitar o acesso intuitivo e dinâmico aos resultados do modelo de predição, permitindo que usuários explorem os dados de forma visual e tabular.

A arquitetura da ferramenta integra as seguintes tecnologias e bibliotecas: ***Streamlit***, para construção da interface web e gestão da interação com o usuário; ***Pandas***, para manipulação e filtragem dos dados agrícolas e climáticos pré-processados; ***Plotly***, para geração dos mapas coropléticos dinâmicos e gráficos interativos; e ***Shapely*** e ***GeoJSON***, para o tratamento e representação das geometrias espaciais dos municípios da Região Geográfica Intermediária de Santa Maria.

O sistema permite ao usuário selecionar parâmetros como a cultura agrícola, o ano da safra, a variável de interesse, o modelo climático, o tipo de mapa (quantitativo ou percentual) e uma cidade específica ou toda a região. Com base nessas entradas, a aplicação carrega os dados correspondentes de projeções agrícolas e climáticas previamente processados, e apresenta-os em mapas e tabelas interativas, permitindo que tanto as tabelas quanto os mapas possam ser baixados. Esse desenvolvimento metodológico garante não só a validação e exploração dos resultados do modelo preditivo, mas também o fornecimento de uma ferramenta de apoio à tomada de decisão para produtores rurais, técnicos e pesquisadores, potencializando a aplicabilidade prática deste trabalho.

5 RESULTADOS

Neste capítulo são apresentados os resultados obtidos com este trabalho, que são apresentados nas próximas seções. A seção Desempenho dos Modelos Random Forest e XGBoost apresenta os resultados das métricas de desempenho dos modelos propostos. O melhor modelo para cada variável alvo foi utilizado para verificar a relevância de cada variável nos cultivos abordados neste trabalho. Na seção Desempenho dos Modelos LSTM, serão apresentadas as métricas de desempenho desses modelos, novamente, o melhor modelo foi utilizado, mas tendo como objetivo fazer as previsões das próximas 10 safras na região. A importância das variáveis climáticas tem seus resultados apresentados na seção seguinte, contendo gráficos que mostram as médias ao longo das últimas décadas em comparação com as previsões de cada cenário SSP. Os resultados das previsões para as próximas safras na região serão abordados na penúltima seção deste capítulo, por fim, será apresentada a ferramenta de visualização em mapas desenvolvida para melhor compreensão visual e permitindo a visualização dos resultados de cada município pertencente a região abordada.

5.1 DESEMPENHO DOS MODELOS RANDOM FOREST E XGBOOST

Inicialmente, serão apresentados os resultados das métricas de desempenho do algoritmo *Random Forest* e, posteriormente, do *XGBoost* em relação as séries temporais de 20 e 30 anos nos cultivos de arroz e soja na região estudada, os números apresentados pelos 3 modelos de *Random Forest* sobre o cultivo de arroz nos datasets históricos podem ser observados na Tabela 4.

Tabela 4 – Desempenho dos Modelos Random Forest para Rendimento Médio e Quantidade Produzida de Arroz

Série	Modelo	Rendimento Médio do Arroz			Quantidade Produzida de Arroz		
		RMSE	R ²	MAPE	RMSE	R ²	MAPE
30 anos	RF1	1626.95	0.2272	33.04%	26149.43	0.6415	29025.68%
	RF2	1499.95	0.3432	20.99%	15104.51	0.8464	871.96%
	RF3	1602.33	0.2505	21.80%	15104.51	0.8804	871.96%
20 anos	RF1	1489.63	0.0090	23.43%	33771.74	0.4034	27238.83%
	RF2	1554.34	-0.0790	19.75%	17285.45	0.8437	529.11%
	RF3	1562.19	-0.0899	19.83%	17196.43	0.8453	651.80%

Podemos observar que em relação a variável alvo rendimento médio, os modelos não apresentaram bom desempenho no coeficiente de determinação, ou seja, não são

capazes de explicar a variabilidade do rendimento médio do arroz na região, variando entre -0.0899 e 0.3432. Esses resultados são inviáveis para compreender a relação entre o clima e a produtividade, especialmente nas séries de 20 anos, em que os valores negativos sugerem que os modelos performam pior do que uma simples média histórica. Mesmo nas melhores situações (como o modelo RF2 na série de 30 anos, com $R^2 = 0.3432$), a capacidade explicativa ainda é considerada baixa. Os valores de RMSE permanecem altos, em torno de 1500 Kg/ha, e os valores do MAPE também são elevados, chegando a mais de 30% no pior caso (RF1 com 30 anos). Isso indica que os modelos cometem erros graves ao estimar o rendimento médio, limitando a confiabilidade das relações entre clima e produção.

No caso da quantidade produzida, embora alguns modelos apresentem R^2 acima de 0.8, como o RF3 na série de 30 anos, os valores de MAPE são extremamente elevados, com o menor valor sendo equivalente a 529.11%, enquanto o maior chega a 29025.68%. Isso evidencia que, apesar de uma boa explicação estatística da variabilidade (alto R^2), os modelos apresentam grandes erros percentuais em relação aos valores reais, o que compromete seriamente sua utilidade. Tais diferenças podem estar relacionadas à magnitude dos valores envolvidos, visto que enquanto alguns municípios produzem cerca de 5 mil toneladas de arroz, outros produzem mais de 200 mil toneladas, demonstrando uma significativa heterogeneidade nos dados da variável alvo, ainda podem estar relacionados a problemas no escalonamento e pré-processamento dos dados de entrada. A seguir, Tabela 5 apresenta o desempenho dos modelos RF1, RF2 e RF3 no cultivo de soja.

Tabela 5 – Desempenho dos Modelos Random Forest para Rendimento Médio e Quantidade Produzida de Soja

Série	Modelo	Rendimento Médio do Soja			Quantidade Produzida de Soja		
		RMSE	R^2	MAPE	RMSE	R^2	MAPE
30 anos	RF1	681.63	0.4945	40.27%	45261.13	0.4125	498.25%
	RF2	685.08	0.4894	35.36%	51738.03	0.2323	65.64%
	RF3	681.10	0.4953	34.89%	50923.43	0.2563	66.11%
20 anos	RF1	653.78	0.5165	52.20%	43553.06	0.2992	830.42%
	RF2	616.59	0.5700	42.01%	31895.17	0.6241	61.51%
	RF3	616.59	0.5700	42.01%	36459.90	0.5088	88.63%

De maneira geral, observa-se uma leve melhora em relação a tabela anterior, com valores mais elevados do coeficiente de determinação e reduções significativas em algumas métricas de erro, embora o desempenho ainda permaneça longe do ideal para aplicações preditivas confiáveis.

Para o rendimento médio da soja, os valores do R^2 variam entre 0.4894 e 0.5700, indicando que os modelos conseguem explicar cerca de 49% a 57% da variabilidade, um aumento considerável em relação ao desempenho no cultivo de arroz, apresentado anteri-

ormente. Os valores do MAPE, no entanto, continuam elevados, variando entre 34,89% e 52,20%, o que sinaliza dificuldades dos modelos em fornecer previsões com precisão aceitável. O RMSE também permanece alto, com valores acima de 600 Kg/ha, reforçando que, mesmo explicando parte da variabilidade, os modelos ainda erram de forma significativa na escala absoluta.

Quanto à quantidade produzida de soja, observa-se uma maior oscilação nos desempenhos. Apesar de alguns modelos apresentarem coeficiente de determinação mais robustos, como o RF2 com 0.6241 (20 anos), os valores de MAPE ainda são bastante elevados, superando 60% em todos os casos e atingindo picos como 830,42% no RF1 (20 anos), o que compromete seriamente a confiabilidade das estimativas. A disparidade entre o R^2 e o MAPE pode indicar que, embora o modelo capture razoavelmente a tendência geral dos dados, ele falha em fornecer previsões precisas, possivelmente devido à presença de valores extremos.

Portanto, embora os resultados desta configuração apresentem melhora em relação às tentativas anteriores, as métricas ainda demonstram que os modelos *Random Forest*, mesmo com séries históricas mais longas e variações estruturais, enfrentam limitações consideráveis na previsão acurada da produção de soja. Isso reforça a necessidade de explorar outros algoritmos como o *XGBoost* que tem suas métricas apresentadas a seguir na Tabela 6.

Tabela 6 – Desempenho dos Modelos XGBoost para Rendimento Médio e Quantidade Produzida de Arroz

Série	Modelo	Rendimento Médio do Arroz			Quantidade Produzida de Arroz		
		RMSE	R^2	MAPE	RMSE	R^2	MAPE
30 anos	XGB1	1206.15	0.5753	25.09%	13291.69	0.9074	21902.13%
	XGB2	1185.44	0.5897	13.48%	7814.88	0.9680	124.03%
	XGB3	1309.31	0.4995	15.42%	7153.67	0.9732	286.58%
20 anos	XGB1	1460.03	0.0480	25.07%	13956.43	0.8981	21090.46%
	XGB2	2066.57	-0.9073	29.68%	17553.46	0.8388	3653.29%
	XGB3	1962.55	-0.7202	30.13%	20553.57	0.7790	4780.27%

As métricas de desempenho dos modelos XGBoost aplicados à previsão do rendimento médio e da quantidade produzida de arroz mostram que, para o rendimento médio com 20 anos de série temporal, nenhum modelo apresentou desempenho satisfatório: os valores de R^2 foram próximos de zero ou negativos e os erros percentuais (MAPE) superaram 25%. Esse resultado pode estar relacionado à baixa quantidade de dados, já que poucas cidades passaram pelo filtro de terem produzido mais de 10 safras no período, comprometendo o aprendizado do modelo. Já com a série de 30 anos, os resultados melhoraram, com destaque para o modelo XGB2, que obteve o menor RMSE (1185.44 kg/ha) e MAPE (13.48%) e o maior R^2 , 0.5897, sendo o mais eficiente para essa variável.

Para a quantidade produzida, todos os modelos com 30 anos apresentaram altos

R^2 , acima de 0.90, indicando boa capacidade de explicação. No entanto, os valores de MAPE foram excessivamente altos em todos os casos, superando 100% e chegando a 21902.13%, o que evidencia erros absolutos elevados, possivelmente causados por valores extremos. Entre os modelos testados, o XGB2 se destaca como o melhor, por apresentar o melhor equilíbrio entre explicabilidade e erro nas duas variáveis. A Tabela 7 apresenta as métricas dos modelos XGB no cultivo da soja.

Tabela 7 – Desempenho dos Modelos XGBoost para Rendimento Médio e Quantidade Produzida de Soja

Série	Modelo	Rendimento Médio do Soja			Quantidade Produzida de Soja		
		RMSE	R^2	MAPE	RMSE	R^2	MAPE
30 anos	XGB1	674.25	0.5054	45.53%	37286.80	0.6013	445.67%
	XGB2	592.45	0.6181	36.09%	33636.93	0.6755	69.48%
	XGB3	597.12	0.6121	35.41%	34223.04	0.6641	71.80%
20 anos	XGB1	700.62	0.4448	55.70%	26477.79	0.7410	333.36%
	XGB2	614.38	0.5731	44.31%	18729.83	0.8704	74.54%
	XGB3	596.47	0.5976	42.62%	24528.12	0.7777	135.03%

Os modelos de XGBoost apresentaram melhores resultados para o rendimento médio com série de 30 anos, enquanto os modelos com séries 20 anos foram ligeiramente superiores na previsão da quantidade produzida, evidenciando uma sensibilidade do modelo ao volume de dados disponível para cada variável.

Para o rendimento médio da soja, o melhor desempenho foi obtido pelo modelo XGB2 com 30 anos, que alcançou o maior R^2 (0.6181), menor RMSE (592.45 kg/ha) e MAPE de 36.09%. O modelo XGB3 apresentou desempenho semelhante, mas com leve inferioridade nas métricas. Já com 20 anos, os resultados foram consistentemente piores, com maior erro percentual e menor capacidade explicativa, reforçando que séries mais longas contribuem para previsões mais precisas nesta variável.

Quanto à quantidade produzida, o modelo XGB2 com 20 anos se destacou com o melhor R^2 (0.8704) e menor MAPE (74.54%), além do menor RMSE (18729.83), sugerindo que a redução da série histórica não comprometeu a previsão dessa variável tanto quanto no rendimento. Os demais modelos apresentaram erros percentuais muito mais elevados, especialmente o XGB1 com 30 anos (445.67%).

Considerando o conjunto das métricas, o modelo XGB2 é o melhor entre os avaliados, pois apresentou o melhor equilíbrio entre explicabilidade (R^2) e erro (RMSE e MAPE) tanto para o rendimento médio quanto para a quantidade produzida, com desempenho consistente em ambas as séries temporais. Dessa forma o modelo XGB2 será utilizado para compreender a importância das variáveis climáticas para os dois cultivos, soja e arroz na Região Geográfica Intermediária de Santa Maria/RS.

5.2 DESEMPENHO DOS MODELOS LSTM

Nesta seção serão apresentadas as métricas de desempenho para os modelo LSTM e também uma análise de seus resultados, justificando a escolha de determinados modelos para realização das previsões para as variáveis alvo rendimento médio e quantidade produzida de cada cultivo estudado nessa pesquisa. A Tabela 8 apresenta as métricas de desempenho para o cultivo de arroz na região.

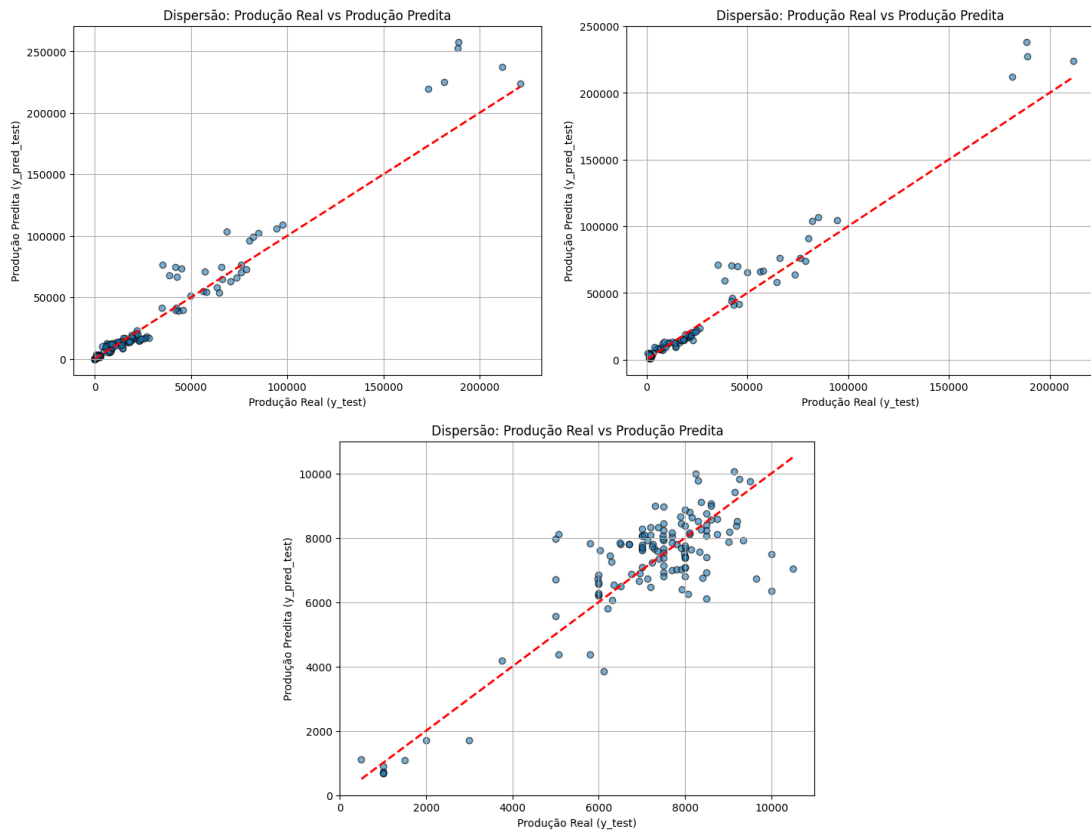
Tabela 8 – Desempenho dos Modelos LSTM para Rendimento Médio e Quantidade Produzida de Arroz

Série	Modelo	Rendimento Médio do Arroz			Quantidade Produzida de Arroz		
		RMSE	R ²	MAPE	RMSE	R ²	MAPE
30 anos	LSTM1	1068.7485	0.6665	12.90%	13100.8223	0.9100	64.51%
	LSTM2	1124.5877	0.6308	18.27%	13786.89	0.4545	67.89%
	LSTM3	1057.1628	0.0813	11.53%	18496.526	0.2412	90.97%
20 anos	LSTM1	938.6203	0.1408	9.47%	11768.0692	0.9311	46.48%
	LSTM2	2457.1628	-0.8731	23.72%	24236.0782	-0.2373	95.72%

A tabela apresenta um desempenho superior dos modelos com série de 30 anos, especialmente no caso do modelo LSTM1, que apresentou resultados expressivos em ambas as variáveis alvo. É importante destacar que o modelo LSTM3 não foi aplicado com os 20 anos, pois já havia tido uma performance ruim com um conjunto de dados maior, a série de 30 anos.

No caso do rendimento médio do arroz, o modelo LSTM1 com 30 anos apresentou o melhor desempenho geral, com RMSE de 1068.75 Kg/ha, MAPE de 12,90% e R² de 0.6665. O modelo LSTM3 apresentou MAPE ainda menor (11,53%), mas com um R² muito baixo (0.0813), indicando que, apesar de errar pouco em média, não conseguiu capturar bem a variabilidade dos dados. Já LSTM2 teve desempenho intermediário. Com 20 anos, o LSTM1 manteve um bom desempenho (MAPE de 9,47%), mas com queda acentuada no coeficiente de determinação (0.1408), mostrando que o modelo se adaptou bem à escala dos valores, mas não conseguiu explicar a variabilidade. Como os modelos não alcançaram um bom desempenho no rendimento médio do arroz com a série de 20 anos, não foram geradas previsões para esta variável com essa série temporal. As previsões do modelo LSTM1 para o conjunto de testes podem ser observadas na Figura 11. O rendimento médio de arroz com a série de 30 anos demonstrou uma distribuição mais esparsa e distante dos valores reais (linha vermelha tracejada), porém, possui uma escala menor em comparação a quantidade produzida, a maioria dos valores preditos, assim como nos valores reais observados, se encontram na faixa com um rendimento entre 6000 Kg/ha e 10.000 Kg/ha.

Figura 11 – Gráficos de Dispersão do conjunto de testes para Quantidade Produzida nas Séries de 30 e 20 anos e Rendimento Médio na Série de 20 anos, respectivamente.



Fonte: Do autor.

Para a quantidade produzida, o modelo LSTM1 novamente se destacou, com R^2 de 0.9100 e MAPE de 64,51% na série de 30 anos. Com 20 anos, seu desempenho foi ainda melhor em termos de R^2 (0.9311) e MAPE mais baixo (46,48%), mostrando que a arquitetura foi capaz de se ajustar bem, mesmo com um dataset reduzido. O modelo LSTM2 teve resultados significativamente piores, com R^2 negativos e altos erros percentuais na série de 20 anos, o que evidencia falhas no aprendizado.

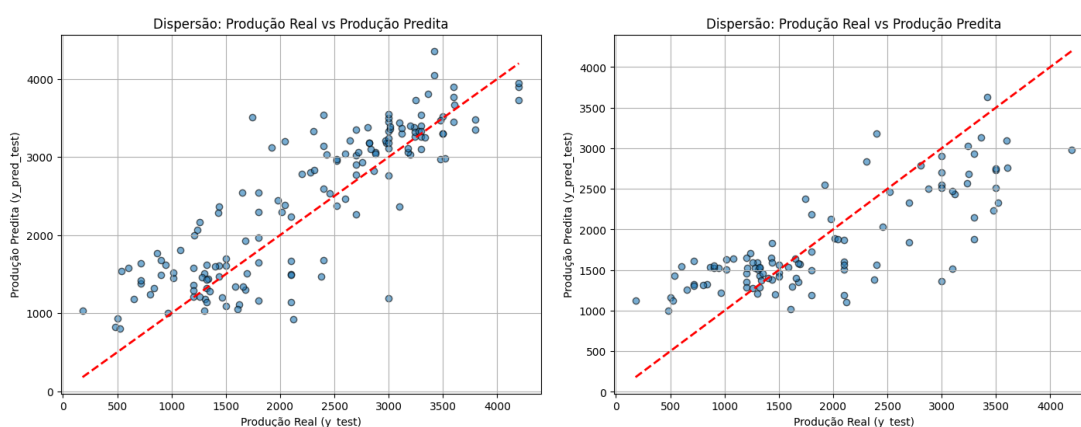
Portanto, o LSTM1 é o melhor modelo entre os testados, por apresentar os melhores resultados de forma consistente tanto para o rendimento médio quanto para a quantidade produzida de arroz. Sua arquitetura mais equilibrada permitiu bom desempenho mesmo com menor volume de dados, portanto para gerar as previsões relacionadas ao cultivo de arroz, foi o modelo utilizado. Em ambos os gráficos de dispersão de quantidade produzida de arroz do modelo LSTM1, pode-se notar as previsões mais próximas da linha tracejada, em comparação ao rendimento médio, no entanto, para valores de produção acima de 50 mil toneladas, as previsões começam a se distanciar dos valores reais, sendo os maiores distanciamentos acima da linha, significando que o modelo realizou previsões acima dos valores reais observados. A seguir, a Tabela 9 apresenta as métricas de desempenho dos modelos LSTM relacionados à soja.

Tabela 9 – Desempenho dos Modelos LSTM para Rendimento Médio e Quantidade Produzida de Soja

Série	Modelo	Rendimento Médio do Soja			Quantidade Produzida de Soja		
		RMSE	R ²	MAPE	RMSE	R ²	MAPE
30 anos	LSTM1	471.7570	0.7579	26.18%	26375.0425	0.8005	46.30%
	LSTM2	503.2563	0.7244	28.08%	29730.9515	0.7465	91.49%
20 anos	LSTM1	593.6214	0.6014	34.57%	20210.6211	0.8491	91.56%
	LSTM2	600.4650	0.5922	36.16%	19178.9517	0.8359	63.09%

Na variável rendimento médio da soja, o melhor desempenho foi alcançado pelo modelo LSTM1 com 30 anos, que apresentou R² de 0.7579, MAPE de 26,18% e o menor RMSE (471.76 Kg/ha). Esse resultado indica uma boa capacidade de explicação da variabilidade dos dados e erro percentual mais controlado em comparação aos demais. Na série de 20 anos, o desempenho geral foi inferior, como esperado, mas o LSTM1 ainda se manteve como o melhor modelo também nesse cenário, com R²=0.6014 e menor MAPE (34,57%), demonstrando maior robustez em ambas as janelas temporais. Os gráficos de dispersão da previsão de rendimento médio do modelo LSTM1 com o conjunto de testes, treinado com as séries de 30 e 20 anos, são apresentados na Figura 12. Em comparação com as previsões da mesma variável alvo no cultivo de arroz, as previsões para a soja se mostraram mais distantes dos valores reais observados e mais distribuídos ao longo do gráfico, enquanto o modelo treinado com a série de 30 anos apresenta um viés mais otimista em relação a maioria das previsões, o modelo treinado com a série de 20 anos, apresentou uma maior quantidade de valores preditos abaixo dos valores reais observados.

Figura 12 – Gráficos de Dispersão do conjunto de testes para Rendimento Médio de Soja nas Séries de 30 e 20 anos, respectivamente.

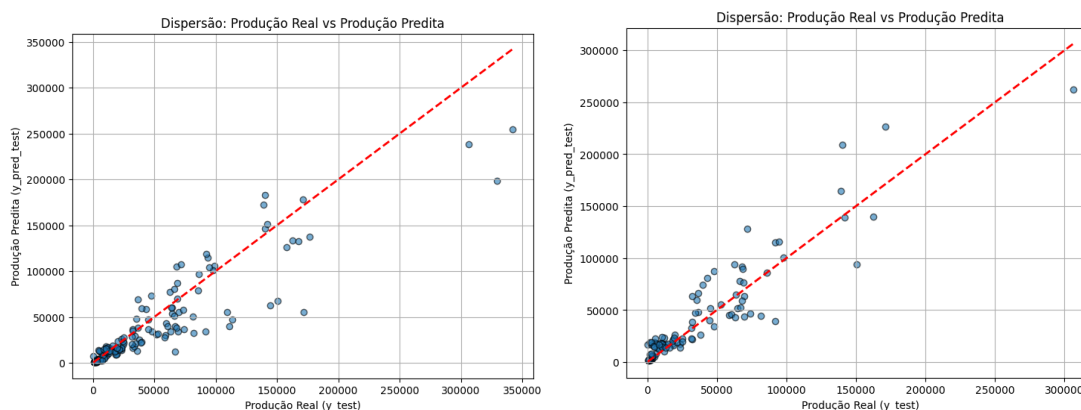


Fonte: Do autor.

Para a quantidade produzida de soja, o melhor resultado na série de 30 anos foi novamente do modelo LSTM1, com R²=0.8005 e MAPE de 46,30%, além do menor RMSE.

No entanto, com 20 anos, o modelo LSTM2 se destacou, apresentando MAPE de 63,09%, bem inferior ao de LSTM1 com 91,56%, com o R^2 ligeiramente menor (0.8359) que o LSTM1, a diferença é mínima, somado ao fato do LSTM2 também possuir um RMSE menor, torna-se a melhor escolha. Isso mostra que, para essa variável, o LSTM2 se adaptou melhor ao menor volume de dados, oferecendo previsões percentuais mais precisas. A Figura 13 apresenta os gráficos de dispersão da quantidade produzida de soja com as previsões do modelo LSTM1 treinado com a série de 30 anos e do modelo LSTM2 com a série de 20 anos no conjunto de testes. Em ambos, pode-se notar uma grande concentração de valores na faixa de 50 mil toneladas produzidas. Na série de 30 anos, as previsões tiveram um grande número de valores abaixo dos valores reais, incluindo os 3 maiores valores observados. Já na série de 20 anos, a distribuição de valores acima e abaixo dos valores reais observados foi mais equilibrada.

Figura 13 – Gráficos de Dispersão do conjunto de testes para Quantidade Produzida de Soja nas Séries de 30 (Modelo LSTM1) e 20 anos (Modelo LSTM2), respectivamente.



Fonte: Do autor.

Pode-se concluir que o modelo LSTM1 com 30 anos é o mais indicado para ambas as variáveis (rendimento médio e quantidade produzida de soja) quando se dispõe de séries históricas mais longas. Já para a série de 20 anos, será usado o LSTM1 para o rendimento médio e o LSTM2 para a quantidade produzida, pois cada um demonstrou desempenho superior em sua respectiva variável alvo. Esses resultados reforçam a importância de ajustar os modelos não apenas à variável predita, mas também à quantidade de dados disponíveis.

5.3 IMPORTÂNCIA DAS VARIÁVEIS CLIMÁTICAS

Nesta seção será apresentada a importância das variáveis climáticas durante as safras de soja e arroz, tendo como variáveis alvo o rendimento médio e quantidade produ-

zida, nas séries temporais de 30 e 20 anos e também nos cenários climáticos abordados nesta pesquisa. Para isso, será utilizado o modelo XGB2 que obteve o melhor desempenho nas métricas de avaliação, como apresentado anteriormente. A seguir, Tabela 10 apresenta a importância relativa das variáveis climáticas na previsão da quantidade produzida e do rendimento médio das culturas de arroz e soja, com base nos dados históricos. A análise considera as duas séries temporais, permitindo identificar quais variáveis mais influenciaram a produção agrícola ao longo do tempo. No Apêndice A, se encontram os gráficos que apresentam as médias de cada variável climática utilizada neste trabalho ao longo dos meses de safra, tanto para as séries temporais quanto as previsões de cada cenário SSP.

Nota-se que não há uma única variável dominante em todos os cenários, a importância relativa varia conforme a cultura, o indicador analisado e o período histórico considerado. No caso do arroz, percebe-se que a umidade do solo tende a ganhar maior importância nas séries de 20 anos, principalmente para o rendimento médio, indicando que em janelas temporais menores o solo atua como um indicador mais estável das condições hídricas acumuladas. Já a temperatura e a precipitação são mais relevantes em séries mais longas, refletindo seu papel na definição do ciclo da planta ao longo do tempo.

Para a soja, a temperatura e a precipitação se destacam na quantidade produzida com 30 anos de dados, o que reforça a sensibilidade da cultura ao regime térmico e hídrico em ciclos completos. Porém, na série de 20 anos, a radiação solar e a umidade do solo passam a ter maior influência, o que pode indicar que, em janelas mais curtas, essas variáveis explicam melhor as flutuações de produtividade. É possível, ainda, que o modelo tenha interpretado a precipitação e a umidade do solo como variáveis com efeitos semelhantes, dado o comportamento inversamente proporcional de suas importâncias em alguns cenários.

Tabela 10 – Importância das Variáveis Climáticas nos Dados Históricos

Variável Alvo	Série Temporal	Precipitação	Temperatura	Radiação Solar	Umidade do Solo
Quantidade Produzida de Arroz	30 anos	17.4%	33.1%	21.8%	27.7%
	20 anos	18.9%	21%	20.9%	39.2%
Rendimento Médio de Arroz	30 anos	27.8%	25.7%	26.7%	19.8%
	20 anos	20.5%	16.5%	18%	45%
Quantidade Produzida de Soja	30 anos	28.5%	35.1%	18.7%	17.7%
	20 anos	23.2%	22.7%	29.2%	24.8%
Rendimento Médio da Soja	30 anos	37.1%	15.2%	31.7%	15.2%
	20 anos	16.9%	26.8%	43.4%	12.9%

A Tabela 11 apresenta a importância das variáveis climáticas na previsão da quantidade produzida e do rendimento médio do arroz, sob os diferentes cenários climáticos abordados (SSP 126, SSP 245, SSP 370 e SSP 585), nas séries temporais de 30 e 20 anos. De modo geral, precipitação e temperatura são as variáveis com maior influência na previsão da quantidade produzida, com destaque para a temperatura no SSP 126 com 20 anos (44,3%) e para a precipitação no SSP 585 com 20 anos (38,7%). A radiação solar aparece com importância intermediária e a umidade do solo tende a ter menor impacto relativo, especialmente em séries mais curtas.

Para o rendimento médio do arroz, avaliado apenas com 30 anos de dados, a precipitação novamente se destaca, sendo a mais relevante em todos os cenários, com destaque para o SSP 370 (46,5%) e SSP 126 (41,4%). A radiação solar também apresenta alta influência, especialmente nos cenários SSP 245 (36,9%) e SSP 585 (36,5%). A temperatura e a umidade do solo mantêm contribuições mais modestas, geralmente abaixo de 22%.

Pode-se concluir que a precipitação foi a variável mais determinante para o cultivo de arroz, principalmente no rendimento médio e em cenários mais extremos. A temperatura ganha relevância em cenários com menos dados, enquanto a radiação solar tem influência consistente como variável complementar.

Tabela 11 – Importância das Variáveis Climáticas para o Cultivo de Arroz

Variável Alvo	Cenário	Série Temporal	Precipitação	Temperatura	Radiação Solar	Umidade do Solo
Quantidade Produzida de Arroz	SSP126	30 anos	28.3%	28.1%	19.6%	24%
		20 anos	23.9%	44.3%	15.7%	16.1%
	SSP245	30 anos	27.8%	22.7%	27.1%	22.4%
		20 anos	31.9%	30%	21.4%	16.7%
	SSP370	30 anos	31.4%	31.5%	16.5%	20.7%
		20 anos	28.6%	37.9%	17.6%	15.8%
	SSP585	30 anos	25.5%	27.2%	25.9%	21.4%
		20 anos	38.7%	30.5%	18.3%	12.5%
Rendimento Médio da Arroz	SSP126	30 anos	41.4%	13.5%	31.2%	13.9%
	SSP245	30 anos	36%	13%	36.9%	14.1%
	SSP370	30 anos	46.5%	21.6%	16.9%	15%
	SSP585	30 anos	32.6%	16.6%	36.5%	14.3%

Os dados apresentados na Tabela 12 tornam visível que nas previsões para a quantidade produzida, há uma distribuição equilibrada entre precipitação, temperatura, radiação solar e umidade do solo, variando conforme o cenário e a série. No SSP 126 com 30 anos, todas as variáveis têm pesos semelhantes, indicando que o modelo considerou múltiplos fatores relevantes para a produção total. Com 20 anos, observa-se um leve aumento da importância da umidade do solo (31,3%), sugerindo sua utilidade compensatória quando há menos dados históricos disponíveis.

A umidade do solo se destaca como a variável mais importante na maioria dos cenários, com relação ao rendimento médio da soja, com destaque para o SSP 370 (64,9% com a série de 30 anos e 71,3% com a série de 20) e SSP 245 (55,9% com 30 anos). Isso pode indicar que a umidade acumulada no solo tem forte relação com a produtividade por hectare, sobretudo em cenários mais quentes e instáveis. Um aspecto interessante é que, em muitos casos, a precipitação apresenta baixa importância, o que pode indicar que o modelo interpretou umidade do solo e precipitação como variáveis fortemente correlacionadas, tratando-as como substitutas em termos de influência.

Desse modo, para a soja, a previsão da quantidade produzida exige uma combinação equilibrada das variáveis climáticas, enquanto o rendimento médio é fortemente influenciado pela umidade do solo, especialmente quando há menos dados disponíveis ou maior variabilidade climática. Essa distinção mostra que diferentes variáveis se desta-

cam conforme a natureza do alvo predito, reforçando a importância de ajustar os modelos conforme o tipo de previsão.

Tabela 12 – Importância das Variáveis Climáticas para o Cultivo de Soja

Variável Alvo	Cenário	Série Temporal	Precipitação	Temperatura	Radiação Solar	Umidade do Solo
Quantidade Produzida de Soja	SSP126	30 anos	20.7%	28.6%	25.7%	25%
		20 anos	24.7%	16.5%	27.5%	31.3%
	SSP245	30 anos	24.3%	30.4%	30.3%	15%
		20 anos	21.4%	25.5%	25.3%	27.8%
	SSP370	30 anos	30.7%	31.7%	14.8%	22.8%
		20 anos	27.5%	26.2%	17.1%	29.2%
	SSP585	30 anos	25.2%	31.4%	26.3%	17.1%
		20 anos	25.6%	24%	29.2%	21.2%
Rendimento Médio da Soja	SSP126	30 anos	22.1%	17.9%	13.1%	46.9%
		20 anos	24.9%	13.5%	9.8%	51.9%
	SSP245	30 anos	9.3%	14.8%	20%	55.9%
		20 anos	8.3%	22.1%	23.2%	46.4%
	SSP370	30 anos	8.5%	12.3%	14.3%	64.9%
		20 anos	8.6%	9.2%	10.9%	71.3%
	SSP585	30 anos	12.4%	36.5%	32.8%	18.2%
		20 anos	16.7%	29.7%	10.5%	43.1

5.4 PREVISÕES PARA AS PRÓXIMAS 10 SAFRAS DE SOJA E ARROZ

Dando continuidade a análise, esta seção apresenta as previsões relacionadas ao cultivo de arroz e soja na Região Geográfica Intermediária de Santa Maria/RS, com foco em duas variáveis fundamentais: a quantidade produzida e o rendimento médio. Inicialmente, são discutidos os dados históricos dessas culturas, permitindo compreender os padrões observados nas últimas décadas. Em seguida, são exibidas as projeções futuras obtidas a partir de modelos treinados com séries temporais de 30 e 20 anos, considerando os diferentes cenários climáticos do CMIP6 (SSP 126, SSP 245, SSP 370 e SSP 585). Para ambas as culturas, será possível avaliar se as variações na produção estão mais re-

lacionadas à área colhida ou ao rendimento por hectare, bem como identificar possíveis vieses nos modelos devido às tendências recentes. Por fim, as análises discutem os resultados de forma crítica, destacando padrões, inconsistências e implicações práticas para o setor agrícola da região.

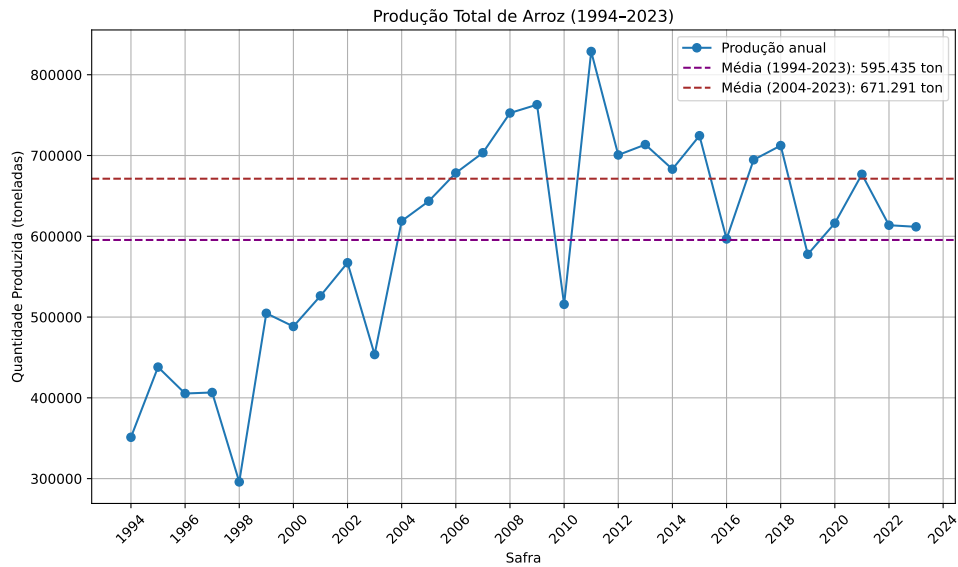
5.4.1 Previsões para o cultivo de Arroz

A quantidade produzida de arroz na Região Geográfica Intermediária de Santa Maria/RS vem se mantendo acima de 500 mil toneladas desde a safra de 2004. Desde então, apenas duas safras (2010 e 2019) registraram produção abaixo de 600 mil toneladas, o que aumenta a média da quantidade produzida na série temporal de 20 anos, 671.291 toneladas, contra 595.435 toneladas de arroz na série de 30 anos conforme mostrado na Figura 14. É perceptível o aumento na produção, vale ressaltar que a quantidade produzida na região é equivalente a soma da produção de todos os municípios pertencentes a ela. Há uma relação entre as variáveis rendimento médio e quantidade produzida, conforme mostrado na equação a seguir:

$$\text{Rendimento Médio} = \frac{\text{Quantidade Produzida(Kg)}}{\text{Area Colhida(ha)}} \quad (5.1)$$

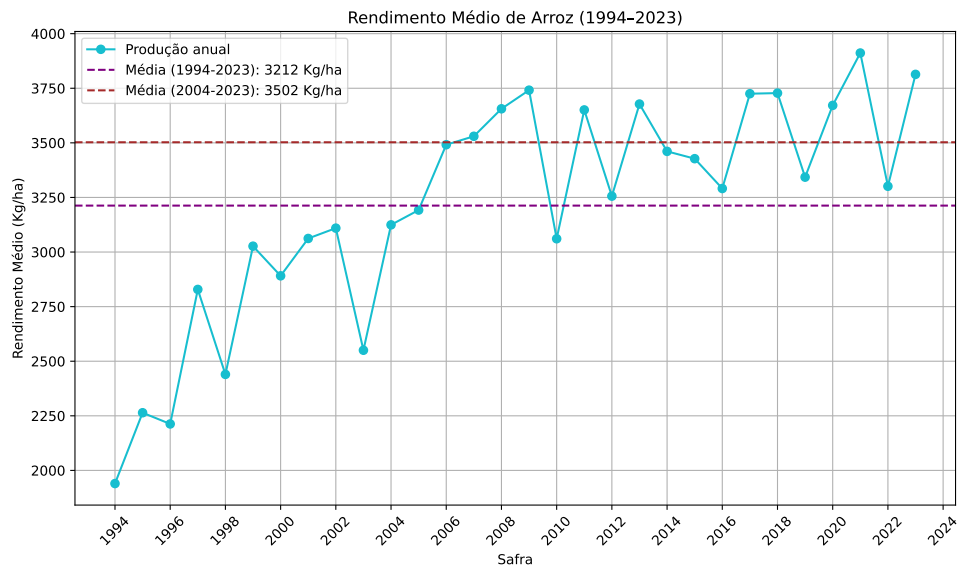
Esta relação é fundamental pra interpretar se o motivo do aumento na produção total se deve a uma maior área colhida, ou seja, a produção aumentou graças a destinação de mais terras ao cultivo do arroz, ou então, foi causada por um aumento da produção por hectare, o rendimento médio. Como apresentado na Figura 15 podemos notar que o ganho na produção está relacionado principalmente a um maior rendimento médio que se manteve acima da média de 30 anos (3212 Kg/ha) desde a safra de 2011, aumentando a média da série de 20 anos para 3502 Kg/ha, mesmo com uma produção total em relativa queda após a safra de 2011, o rendimento médio se manteve estável nesse período, o que demonstra um aumento de produtividade no cultivo de arroz na região, dessa forma o crescimento da quantidade produzida não está relacionado necessariamente a um aumento da área colhida.

Figura 14 – Histórico de Quantidade Produzida de Arroz.



Fonte: Do autor.

Figura 15 – Histórico de Rendimento Médio do Arroz.

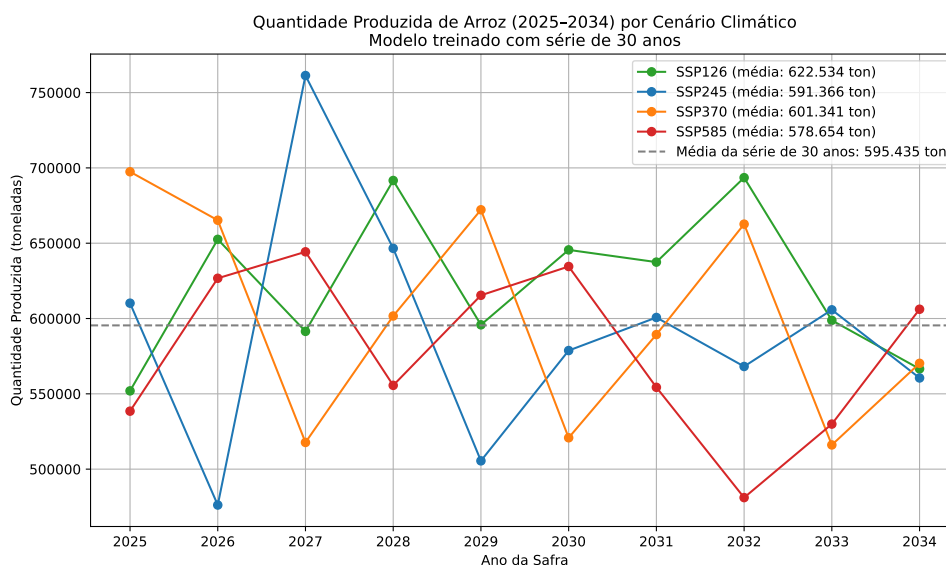


Fonte: Do autor.

A análise do histórico feita anteriormente é de grande importância para compreensão dos resultados de produção total e rendimento médio do arroz na região, apresentados pelos 4 cenários climáticos SSP (126, 245, 370, 585). Com o modelo treinado com a série temporal de 30 anos, a Figura 16 mostra que todos os cenários possuem variações na quantidade produzida de arroz para próximas 10 safras, com valores tendo um certo padrão de alternância com valores acima e abaixo da média histórica de 30 anos (595.654 toneladas) a cada 2 safras. A média de cada cenário entre 2025 e 2034, mostra que a produção total de arroz nesse período segue a ordem de otimismo dos cenários, do melhor

ao pior. O cenário SSP 126, apresentou a maior média, com 622.534 ton, seguido por SSP 370 (601.341 ton), SSP 245 (591.366 ton) e SSP 585 (578.654 ton). O cenário SSP 245 chamou atenção pelo fato de que em duas safras consecutivas (2026 e 2027) demonstrou o menor e maior valor de produção respectivamente, e na safra de 2029 apresentou a 3ª pior safra entre todos os cenários, se tornando o cenário com as maiores oscilações, porém, ainda manteve uma média superior ao cenário climático mais pessimista, SSP 585.

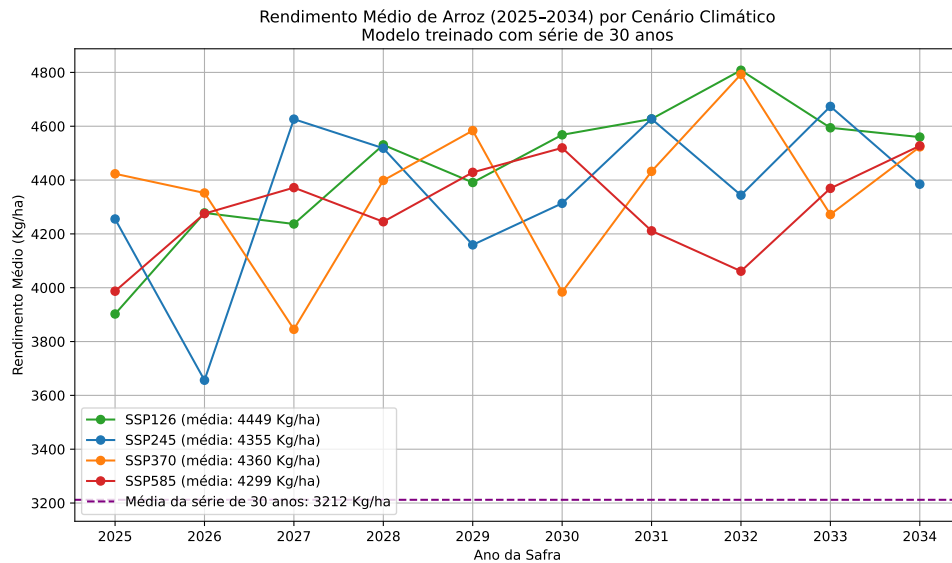
Figura 16 – Previsão Quantidade Produzida de Arroz com Série de 30 anos.



Fonte: Do autor.

Para compreender se as variações da produção de arroz na série de 30 anos estão mais relacionadas a área colhida ou ao rendimento em si, a Figura 17 demonstra um resultado de certa forma surpreendente, os 4 cenários apresentam rendimento bem acima da média histórica da série de 30 anos. A menor previsão de rendimento médio, na safra de 2026 com o cenário SSP 245, é mais de 1000 Kg/ha acima da média de 30 anos, demonstrando um ótimo rendimento, o cenário SSP 126 novamente apresenta a melhor média com 4449 Kg/ha, seguido por SSP 370 com 4360 Kg/ha, SSP 245 apresentando rendimento médio de 4355 Kg/ha e, SSP 585 com 4299 Kg/ha. Portanto, com um rendimento bem acima da média, é possível inferir que as predições de quantidade produzida de arroz abaixo da média histórica em algumas safras, estão relacionadas a uma menor área colhida, podendo ser causadas pelas próprias ações climáticas em determinadas áreas e/ou possíveis decisões dos produtores rurais que podem também optar por outros cultivos em suas terras.

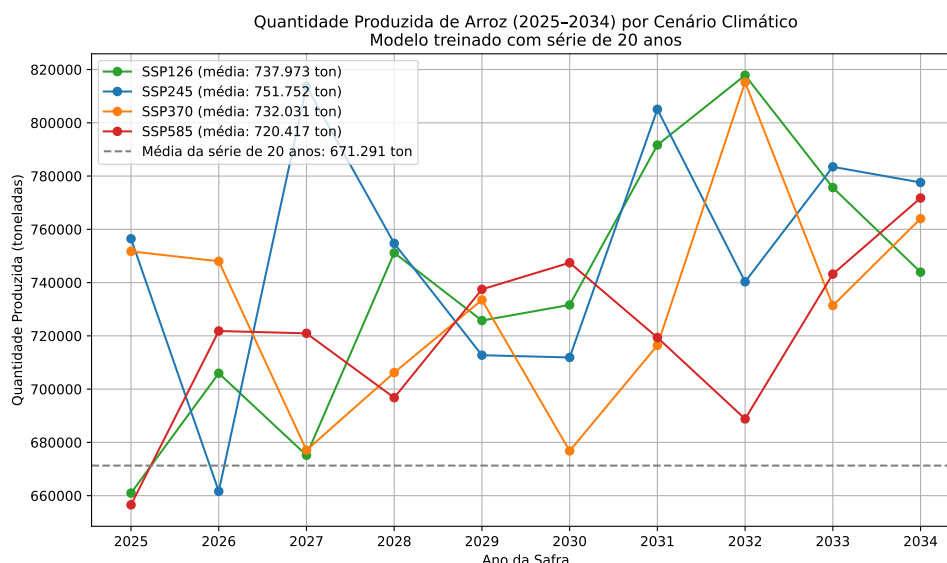
Figura 17 – Previsão Rendimento Médio do Arroz com Série de 30 anos.



Fonte: Do autor.

Nos resultados do modelo treinado com a série de 20 anos, mostrado na Figura 18 pode-se observar que o modelo capturou com mais evidência a tendência de crescimento da produção nos últimos anos, o que já é esperado, devido ao grande aumento apresentado desde a safra de 2011, o que representa mais da metade das safras do modelo de 20 anos. Dessa forma, considerando os 40 pontos no gráfico (4 cenários com 10 anos cada) apenas 3 ficaram abaixo da média histórica de 20 anos, nas safras de 2025 e 2026. Se torna evidente que o modelo apresentou resultados mais otimistas causados pelo viés de crescimento, especialmente da última década. O modelo SSP 245 apresentou a melhor média equivalente a 751.752 ton, seguido pelo SSP 126 (737.973 ton), SS370 (732.031 ton) e SSP 585 (720.417 ton).

Figura 18 – Previsão Quantidade Produzida de Arroz com Série de 20 anos.



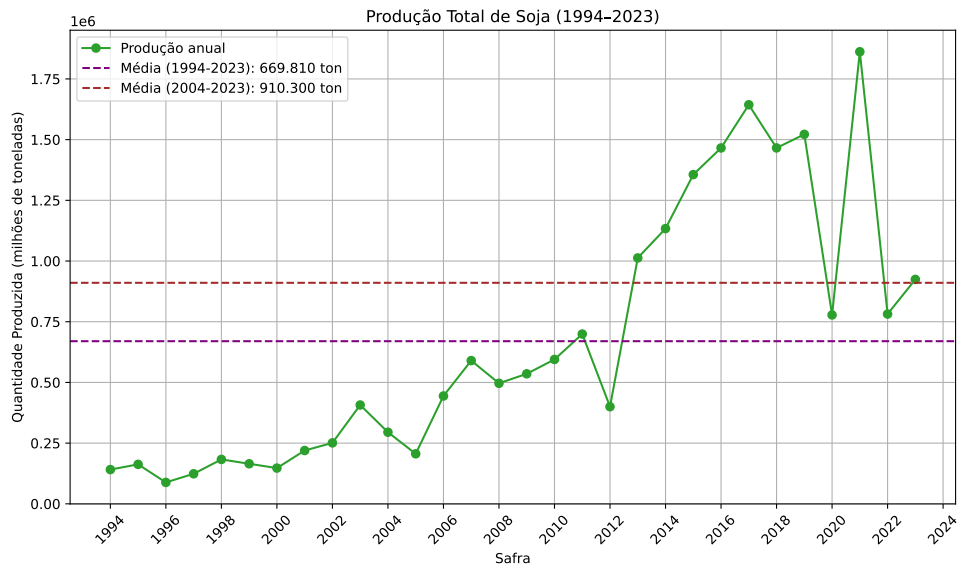
Fonte: Do autor.

5.4.2 Previsões para o cultivo de Soja

O histórico de quantidade produzida de soja na região (Figura 19), demonstra um grande crescimento deste cultivo na região, a partir da safra de 2013, todas apresentaram produção acima da média de 30 anos que é equivalente a 669.810 toneladas. Assim como no cultivo do arroz, esse crescimento na última década impacta na média da série de 20 anos que sobe para 910.300 toneladas. Desde o ano de 2013, apenas duas safras apresentaram resultados abaixo da média de 20 anos, 2020, ano da pandemia da Covid-19 e 2022, ano no qual o estado do Rio Grande do Sul foi fortemente impactado pelo fenômeno *La Niña*, que causou estiagem prolongada e temperaturas acima da média.

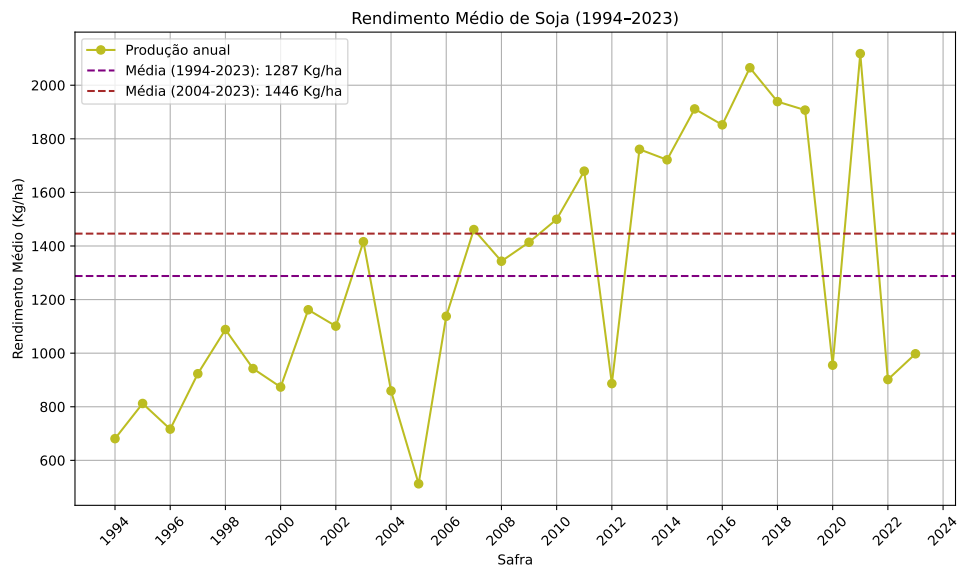
Ao observar a Figura 20, que demonstra o rendimento médio da soja na região, é possível perceber a semelhança com o gráfico de quantidade produzida, nos anos de menor produção, o rendimento médio foi menor, nos anos de maior produção, o rendimento também foi maior. Isso sugere uma relação na qual as condições climáticas, tecnológicas ou de manejo (que afetam o rendimento) estão sendo os fatores mais relevantes, e não a área cultivada. Esse comportamento pode significar que produtores não estão expandindo a área (por limite físico, ambiental ou econômico), mas em produtividade, ou seja: melhoramento genético, fertilização, controle de pragas, irrigação, entre outras ações.

Figura 19 – Histórico de Quantidade Produzida de Soja.



Fonte: Do autor.

Figura 20 – Histórico de Rendimento Médio da Soja.

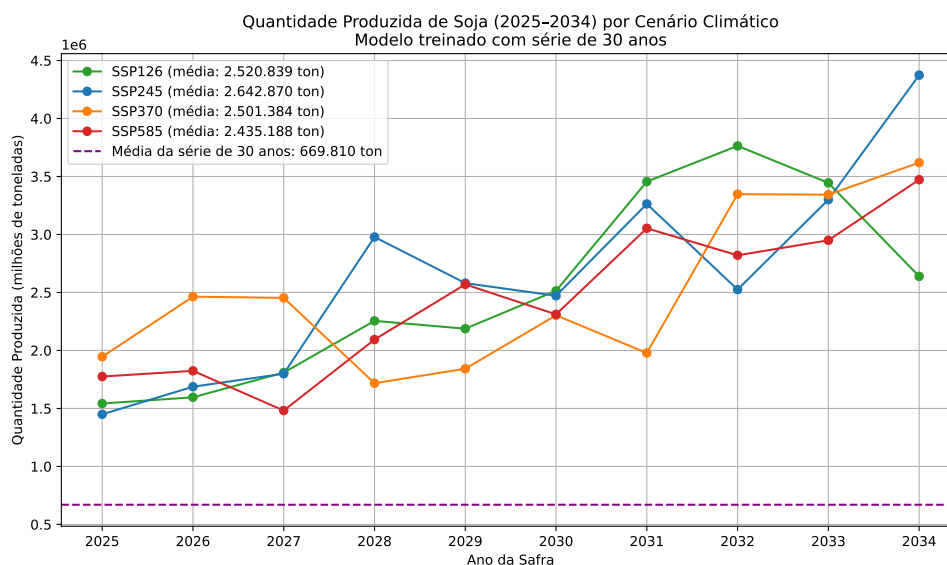


Fonte: Do autor.

O resultado das previsões de quantidade produzida de soja para as próximas 10 safras apresentados na Figura 21 foram extremamente otimistas, as médias de produção em todos os cenários são quase 4 vezes superior a média de 30 anos, o valor do RMSE equivalente a 26375 toneladas pode ter grande influencia nesse resultado, podendo ter superestimado a quantidade produzida de soja. O gráfico também demonstra uma tendência de aumento da produção ao longo dos anos, a cada safra o cenário com a menor produção, foi superior a menor estimativa da safra anterior com exceções das safras de 2031 e 2034. A última safra prevista surpreende com o resultado no cenário SSP 126 que apresentou

o pior desempenho entre os cenários, derrubando sua média de produção no período. O cenário SSP 245 apresentou a melhor média com 2.642.870 ton, seguido pelo SSP 126 com 2.520.839 ton, SSP 370 com média de 2.501.188 ton e SSP 585 com produção total de 2.435.188 ton.

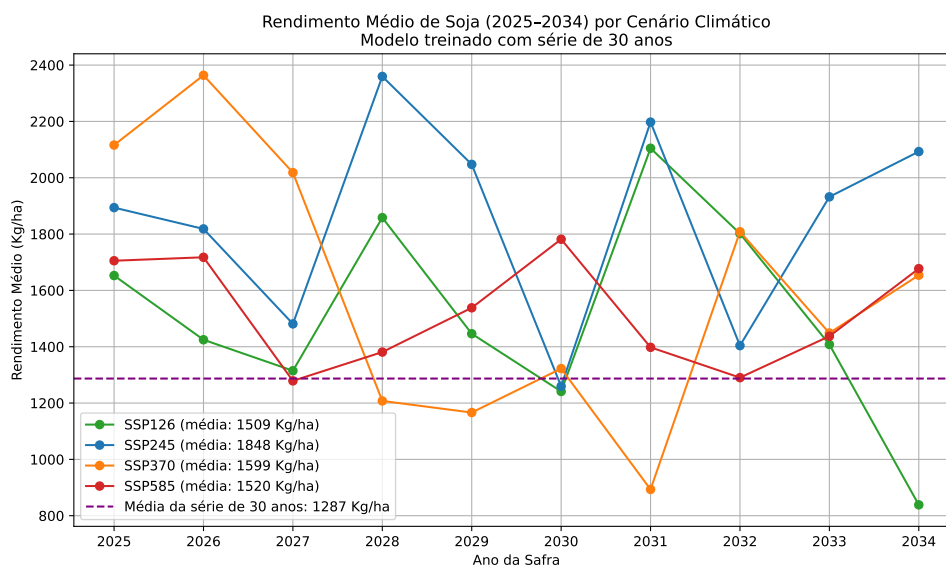
Figura 21 – Previsão Quantidade Produzida de Soja com Série de 30 anos.



Fonte: Do autor.

Os resultados da previsão de rendimento médio da soja com a série temporal de 30 anos, demonstrados pela Figura 22 permite melhores conclusões sobre a previsão relacionada a quantidade produzida. Todos os cenários tiveram grande alternância entre os resultados ao longo das 10 safras previstas, mesmo com os cenários apresentando rendimento médio acima da média de 30 anos na maioria das safras, o gráfico demonstra uma certa semelhança com o gráfico de quantidade produzida, porém, com um rendimento ainda próximo da média e algumas vezes até inferior, contrastando com a quantidade produzida que alcança valores quase 4 vezes superior a média. Para que isso seja possível, seria necessário aumentar a área colhida de soja na região em quase 4 vezes também, o que se torna inviável devido a questões físicas, ambientais e econômicas. O cenário com melhor média de rendimento foi o SSP 245 (1848 Kg/ha), seguido por SSP379 (1599 Kg/ha), SSP 585 (1520 Kg/ha) e SSP (1509 Kg/ha).

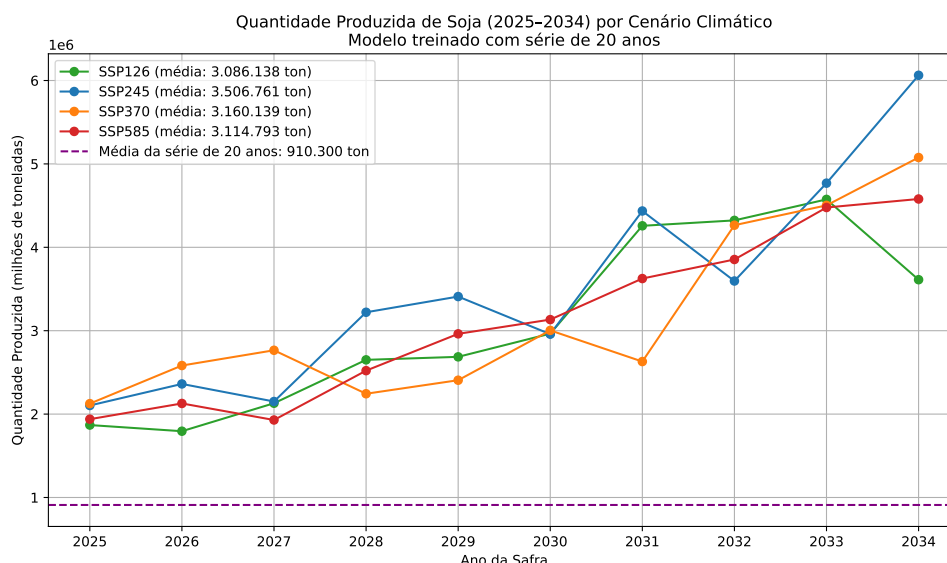
Figura 22 – Previsão Rendimento Médio da Soja com Série de 30 anos.



Fonte: Do autor.

Por fim, serão apresentados os resultados das previsões com modelos treinados com séries de 20 anos para o cultivo do soja na Região Geográfica Intermediária de Santa Maria/RS. A Figura 23 continua apresentando valores elevados na quantidade produzida de soja na região, assim como ocorreu na mesma variável alvo com o modelo treinado com série de 30 anos, os gráficos, apesar da diferença entre os valores, apresenta um comportamento semelhante entre as duas séries temporais. Com médias em cada cenário no período da previsão, sendo 3 vezes maior que a média histórica dos últimos 20 anos, o resultado também apresenta crescimento constante, com exceções das safras de 2031 e 2034, o cenário com a pior produção foi superior em relação ao pior da última safra. Os elevados valores de produção de alguns municípios podem ter enviesado a previsão da produção de soja, tanto na série de 30 quanto de 20 anos, mesmo a aplicação da escala logarítmica pode não ter conseguido evitar esta situação. No modelo treinado com série de 20 anos, o cenário SSP 245 apresentou a melhor média no período com produção de 3.506.761 ton, seguido por SSP 370 (3.160.139 ton), SSP 585 (3.114.793 ton) e SSP 126 (3.086.138 ton).

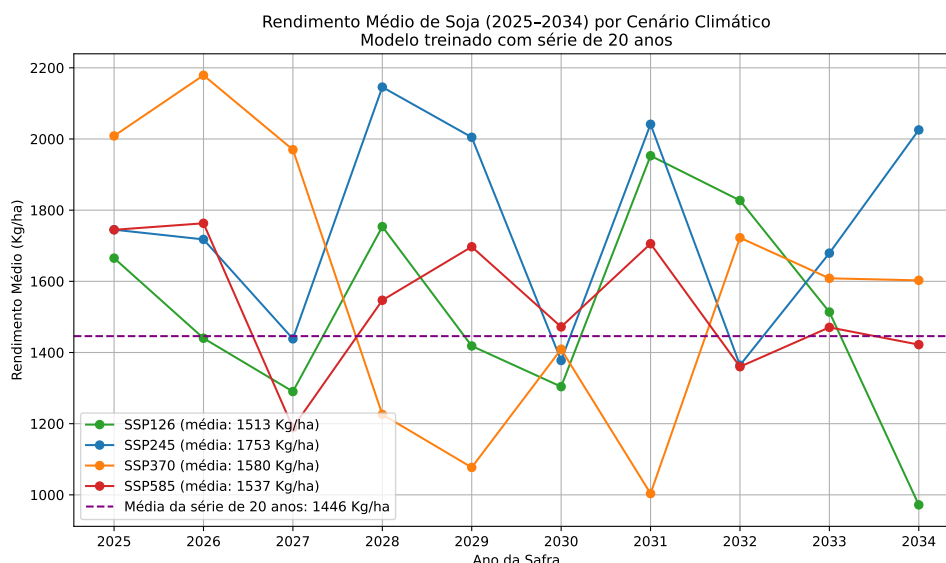
Figura 23 – Previsão Quantidade Produzida de Soja com Série de 20 anos.



Fonte: Do autor.

Os resultados da previsão de rendimento médio da soja na região estudada nesta pesquisa na Figura 24 mostram uma grande variabilidade. Todos os cenários apresentaram ao menos uma safra com rendimento abaixo da média de 20 anos, ao mesmo tempo em que outros apresentam valores bem acima da média na mesma safra, dessa forma pode-se compreender a ação climática de cada cenário ao longo dos anos como fator determinante para esse resultado, principalmente relacionado a umidade do solo, variável de maior importância para o rendimento médio do cultivo. Assim como no resultado da série de 30 anos, os valores de rendimento médio demonstram instabilidade e variando acima e abaixo da média histórica, contrastando com o crescimento quase linear da quantidade produzida de soja, a mesma análise de inviabilidade dos valores de quantidade produzida com a série de 30 anos pode ser aplicada neste caso, reforçando que a grande produção de algumas cidades enviesaram o modelo, tornando suas previsões extremamente otimistas e muito acima das médias da série. Percebe-se que cada cenário apresentou média de rendimento próximo a do período histórico, com o cenário SSP 245 tendo o melhor rendimento com 1753 Kg/ha, seguido por SSP 370 com 1580 Kg/ha, depois SSP 585 com 1537 Kg/ha e por último SSP 126 com rendimento médio de 1513 Kg/ha, enquanto a média histórica de 20 anos é de 1446 Kg/ha.

Figura 24 – Previsão Rendimento Médio da Soja com Série de 20 anos.



Fonte: Do autor.

5.5 SISTEMA INTERATIVO PARA VISUALIZAÇÃO DOS RESULTADOS

Como mencionado no capítulo anterior, foi desenvolvida uma aplicação interativa utilizando a biblioteca *Streamlit*, voltada à visualização de projeções agrícolas para todos os municípios da Região Geográfica Intermediária de Santa Maria/RS, considerando diferentes cenários climáticos futuros. A aplicação permite ao usuário explorar os impactos climáticos previstos sobre as culturas de soja e arroz por meio de mapas coropléticos dinâmicos e tabelas de dados filtráveis.

A ferramenta foi estruturada para fornecer uma interface intuitiva, na qual o usuário pode selecionar o ano da safra, a cultura agrícola, a variável de interesse (rendimento médio ou quantidade produzida), o modelo climático baseado em séries históricas de 20 ou 30 anos, o tipo de mapa desejado (quantitativo ou percentual) e a cidade específica ou toda a região intermediária. Com base nessas seleções, o sistema carrega automaticamente os dados de projeção climática e agrícola provenientes de arquivos CSV processados anteriormente, os quais estão integrados ao modelo de predição.

A visualização espacial é realizada por meio da biblioteca *Plotly*, com uso de mapas coropléticos que representam os valores absolutos ou percentuais da variável selecionada para os quatro cenários climáticos projetados pelo CMIP6 utilizados nesta pesquisa: SSP 126, SSP 245, SSP 370 e SSP 585. Cada cenário é apresentado com um mapa separado, acompanhado de uma descrição resumida acessível por meio de um ícone interativo com *tooltip*.

Um dos principais diferenciais da aplicação está na funcionalidade de comparação direta com as médias históricas. Para isso, a aplicação carrega também os dados médios municipais da cultura e variável selecionadas, calculados previamente para os períodos históricos correspondentes (20 ou 30 anos). Com isso, o sistema calcula dinamicamente a diferença entre o valor projetado e a média histórica, tanto em termos absolutos quanto percentuais. Essas diferenças são apresentadas diretamente no hover de cada cidade no mapa, com destaque visual (verde para valores acima da média, vermelho para abaixo), facilitando a identificação de tendências positivas ou negativas de forma imediata.

Além dos mapas, a aplicação oferece uma tabela interativa contendo os dados numéricos filtrados de acordo com os parâmetros selecionados. Essa tabela apresenta os valores projetados para cada cenário selecionado, os quais o usuário pode selecionar, a tabela, assim como os mapas, permite uma visualização detalhada por município. Outra funcionalidade pé a possibilidade de download dos mapas como imagem e das tabelas como um arquivo CSV. Por fim, a aplicação inclui um rodapé institucional, com informações do autor, orientador, instituição, ano e link para o repositório no GitHub, a aplicação pode ser acessada por meio da seguinte URL: <https://previsaoagricolaregiaosm.streamlit.app>.

Do ponto de vista técnico, a aplicação integra diversas bibliotecas e tecnologias: *Streamlit* para interface e lógica de aplicação, *Pandas* para manipulação de dados, *Plotly* para geração dos gráficos, e *Shapely* e *GeoJSON* para o tratamento e exibição da geometria espacial dos municípios. Essa combinação garante uma plataforma acessível e de fácil manutenção, contribuindo não apenas para a validação dos resultados obtidos ao longo do trabalho, mas também como uma potencial ferramenta de apoio à tomada de decisão no contexto da agricultura regional.

6 CONCLUSÃO

Este trabalho teve como objetivo principal analisar o impacto das mudanças climáticas na produção agrícola de soja e arroz na Região Geográfica Intermediária de Santa Maria (RS), utilizando algoritmos de aprendizado de máquina e séries temporais de 20 e 30 anos. Foram aplicados os modelos Random Forest, XGBoost e LSTM para identificar variáveis climáticas mais relevantes e prever a produtividade futura sob diferentes cenários do CMIP6.

Os resultados obtidos com os modelos Random Forest e XGBoost demonstraram uma maior capacidade para explicar a variabilidade da quantidade produzida em relação ao rendimento médio, especialmente com séries de 30 anos. O modelo XGB3, por exemplo, atingiu R^2 de 0,9732 na previsão da quantidade de arroz, sugerindo que os dados históricos climáticos e produtivos conseguem explicar bem a variação dessa variável. No entanto, os valores de MAPE extremamente elevados, em alguns casos superiores a 20.000%, como foi o caso do modelo XGB1 na série de 30 anos para a quantidade produzida de arroz, indicando que, apesar da boa explicação estatística, as previsões absolutas foram distorcidas, especialmente em municípios com produções muito altas, o que causou enviesamento dos modelos. Já para o rendimento médio, tanto em arroz quanto em soja, os modelos apresentaram desempenho inferior, com R^2 baixos ou até negativos, especialmente em séries de 20 anos. Isso evidencia a dificuldade desses modelos em captar a influência do clima sobre a produtividade unitária, provavelmente devido à ausência de outras variáveis relevantes, como tipo de solo, práticas de manejo ou uso de tecnologias agrícolas.

Os modelos LSTM, por sua vez, mostraram maior capacidade para capturar os padrões temporais e sazonais das culturas, mesmo que seus resultados ainda tenham sido limitados pela variabilidade dos dados. As previsões futuras indicaram, por exemplo, que o cenário SSP245 apresentou os melhores resultados médios de rendimento para a soja, com 1.753 kg/ha na série de 20 anos, superando a média histórica de 1.446 kg/ha, o que sugere que, sob cenários intermediários de mudança climática, pode haver benefícios agrícolas moderados, desde que outros fatores permaneçam constantes.

No entanto, ao observar a quantidade produzida futura de soja, nota-se uma inconsistência: em todos os cenários, os valores projetados foram quase quatro vezes maiores que a média histórica, o que só seria possível com um aumento equivalente na área colhida, algo fisicamente inviável. Esse comportamento reforça a necessidade de incorporar limites realistas ao modelo, seja por meio de variáveis de área plantada ou de validação externa. A importância das variáveis climáticas, mensurada via Random Forest e XGBoost, mostrou uma grande variação no ranking das variáveis climáticas nas séries históricas e nas projeções, mostrando que eventos meteorológicos fora do padrão causar grande influência na produção agrícola.

O desenvolvimento do sistema interativo baseado em mapas facilitou a visualização dos resultados por município e safra, permitindo a exploração dos impactos das mudanças climáticas de forma intuitiva e personalizada. Essa ferramenta representa um diferencial da pesquisa, pois traduz os dados em informações acionáveis, com grande potencial de aplicação prática. Assim, conclui-se que a integração de dados climáticos com aprendizado de máquina oferece uma abordagem promissora para antecipar os impactos das mudanças climáticas na agricultura. Contudo, os resultados também indicam limitações importantes, especialmente na modelagem de rendimento médio e nas previsões absolutas exageradas.

6.1 TRABALHOS FUTUROS

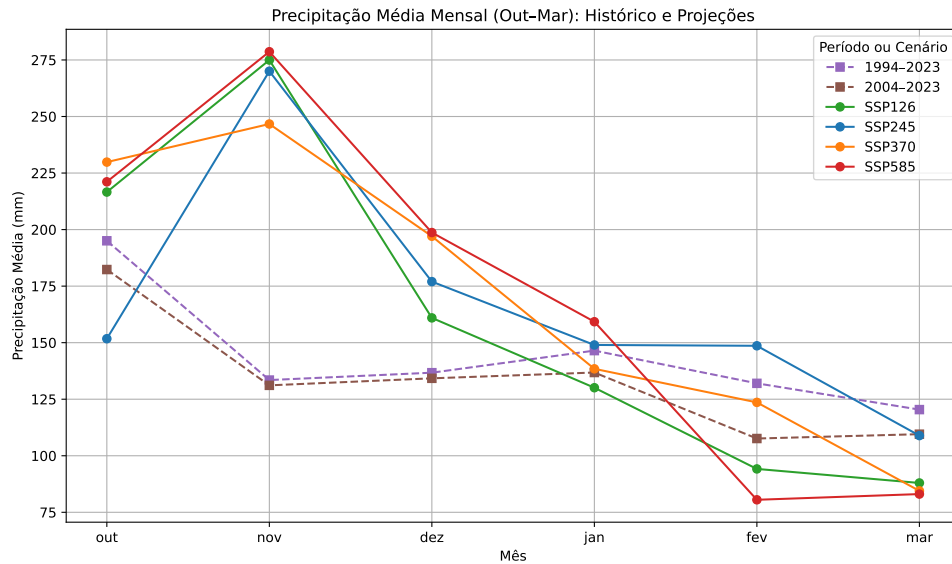
Como sugestão para trabalhos futuros, o uso de um número maior de municípios, abrangendo uma região geográfica mais extensa, pode melhorar o desempenho dos modelos. Isso permitiria a incorporação de uma maior diversidade de dados climáticos e produtivos, tornando as análises mais robustas e generalizáveis. Além disso, a metodologia proposta pode ser aplicada em qualquer outra região do Brasil, o que abre espaço para estudos comparativos entre diferentes contextos agroclimáticos. Também é possível expandir o escopo para abranger outros cultivos além da soja e do arroz, contribuindo para a construção de um sistema mais abrangente e aplicável a uma variedade maior de produtores.

Outra proposta relevante para trabalhos futuros envolve a inclusão de variáveis relacionadas ao solo, como tipo, textura, fertilidade, que possuem grande influência sobre o rendimento agrícola e podem melhorar o desempenho dos modelos preditivos. A integração de dados de previsão meteorológica de alta resolução é outra melhoria promissora, especialmente para análises de curto prazo e apoio à tomada de decisão em tempo real. Por fim, o sistema interativo desenvolvido pode ser aprimorado com a inclusão de gráficos históricos de produção e das variáveis climáticas, bem como a implementação de filtros e opções de visualização mais detalhadas. Essas melhorias aumentariam significativamente a interatividade e a utilidade prática da ferramenta para os usuários.

Além disso, melhorias nos modelos preditivos utilizados também representam uma possibilidade para trabalhos futuros. No caso dos algoritmos Random Forest e XGBoost, é possível explorar formas mais refinadas de ajuste de hiperparâmetros e seleção de variáveis, buscando melhorar a performance e a interpretabilidade dos resultados. Já para o modelo LSTM, futuras abordagens podem considerar ajustes na arquitetura da rede, no tamanho da janela temporal e no tratamento dos dados de entrada, com o objetivo de capturar melhor os padrões temporais e aumentar a precisão das previsões. Essas melhorias podem contribuir para tornar os modelos mais robustos e adaptáveis a diferentes contextos agrícolas.

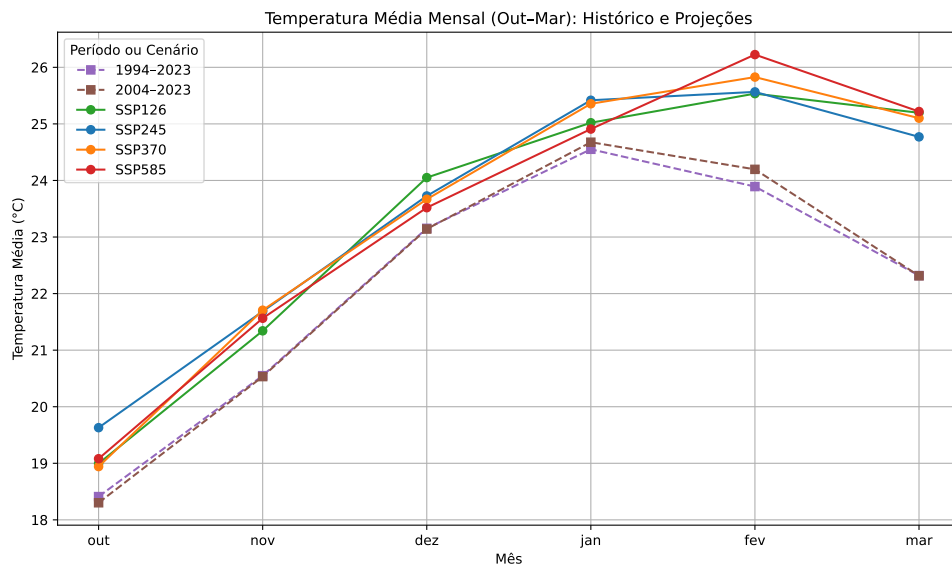
APÊNDICE A – GRÁFICOS COMPLEMENTARES

Figura 25 – Precipitação média nos meses de safra.



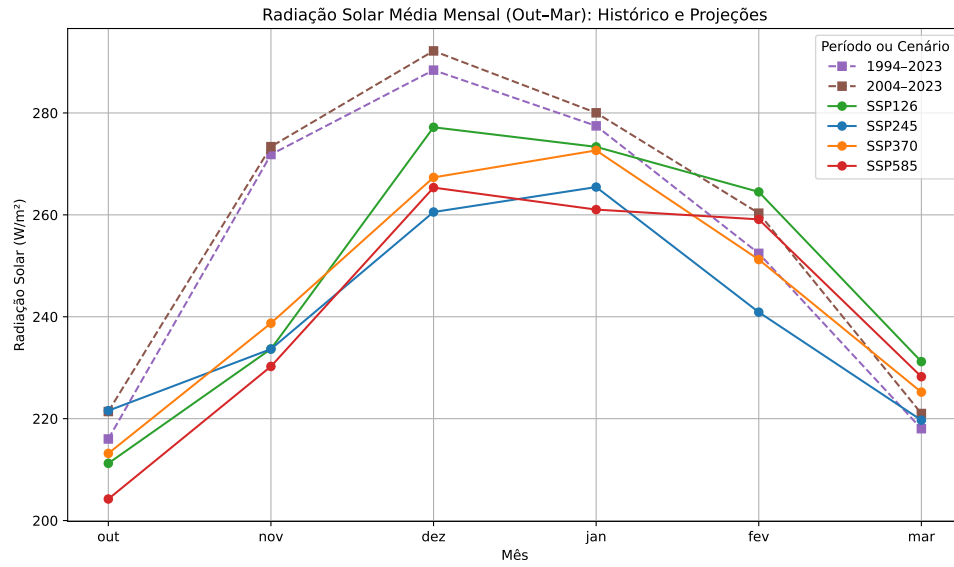
Fonte: Elaborado pelo autor.

Figura 26 – Temperatura média nos meses de safra.



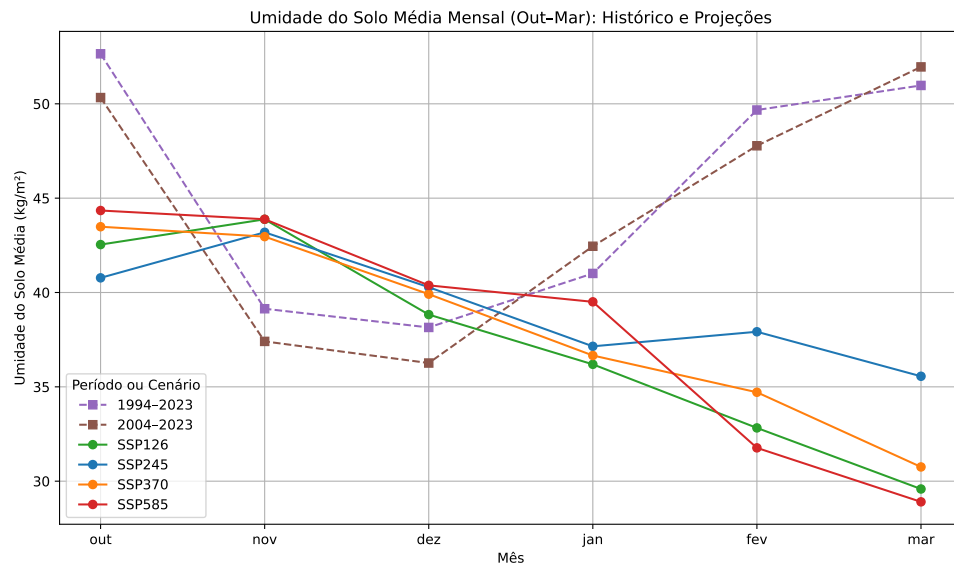
Fonte: Elaborado pelo autor.

Figura 27 – Radiação solar nos meses de safra.



Fonte: Elaborado pelo autor.

Figura 28 – Umidade do solo nos meses de safra.



Fonte: Elaborado pelo autor.

REFERÊNCIAS

- ARSEGO, D. A. **Modelo Estatístico de Previsão de Produtividade de Soja e Arroz para o Rio Grande do Sul**. 2017. 156 p. Tese (Tese de Doutorado) — Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Programa de Pós-Graduação em Meteorologia, 2017.
- BARNI, M. A.; MATZENAUER, R. **Ampliação do calendário de semeadura da soja no Rio Grande do Sul pelo uso de cultivares adaptados aos distintos ambientes**. Porto Alegre, 2014. Acesso em: 25 abr. 2025. Disponível em: <http://www.fepagro.rs.gov.br/upload/1398888806_art_02.pdf>.
- BATISTELLA, D. **Estimativa de Produtividade de Soja por Meio de Imagens Orbitais e Machine Learning**. 2023. Tese (Tese de Doutorado) — Universidade Tecnológica Federal do Paraná, 2023.
- BBC News Brasil. **Rio Grande do Sul ainda vai viver muitos eventos extremos, dizem cientistas brasileiras que colaboraram com IPCC**. 2024. <https://www.bbc.com/portuguese/articles/czkv2mrdv31o>. Acesso em: 28 abr. 2025.
- BERGAMASCHI, H. et al. Distribuição hídrica no período crítico do milho e influência na produção de grãos. **Revista Brasileira de Agrometeorologia**, v. 12, n. 2, p. 343–348, 2004.
- BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis: Forecasting and Control**. 4. ed. Hoboken, NJ: John Wiley & Sons, 2008.
- BREIMAN. Random forests. **Machine Learning**, Springer, v. 45, p. 5–32, 2001.
- BREIMAN, L. Bagging predictors. **Machine learning**, v. 24, p. 123–140, 1996.
- CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. **Geoscientific Model Development**, v. 7, p. 1247–1250, 2014. Disponível em: <<https://doi.org/10.5194/gmd-7-1247-2014>>.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 785–794.
- COUNCE, P. A.; L., N. T.; FEHR, W. R. A uniform, objective, and adaptative system for expressing rice development. **Crop Science**, Madison, v. 40, n. 2, p. 436–443, 2000.
- CUNHA, G. R.; WREGE, M. S.; MALUF, J. R. T. Clima e agricultura no brasil: riscos e oportunidades. In: MONTEIRO, J. E. B. A. (Ed.). **Agrometeorologia dos Cultivos: o fator meteorológico na produção agrícola**. Brasília: INMET, 2001. p. 17–30.
- DOMINGOS, P. **O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo**. [S.l.]: Novatec Editora, 2017. 400 p.
- EMATER/RS-ASCAR. **Boletim Adverso N.º 1: Impactos das chuvas e cheias extremas no Rio Grande do Sul em maio de 2024**. Porto Alegre: Governo do Estado do Rio Grande do Sul, 2024. Coordenação: Gerência de Planejamento GPL (Núcleo de Informações e

Análises NIA). Acesso em: 25 abr. 2025. Disponível em: <<https://www.estado.rs.gov.br/upload/arquivos/202406/relatorio-sisperdas-evento-enchentes-em-maio-2024.pdf>>.

Empresa Brasileira de Assistência Técnica e Extensão Rural; Empresa Brasileira de Pesquisa Agropecuária. **Sistemas de Produção para a Cultura do Arroz Irrigado e de Sequeiro; Zona da Mata - MG**. Brasil, 1981. 28 p. (Sistemas de Produção - Boletim nº 316).

EOS Data Analytics. **Plantação de arroz: guia completo para agricultores**. 2024. Acesso em: 25 abr. 2025. Disponível em: <<https://eos.com/pt/blog/plantacao-de-arroz/>>.

EYRING, V. et al. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. **Geoscientific Model Development**, v. 9, n. 5, p. 1937–1958, 2016.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FEHR, W. R.; CAVINESS, J. D. **Stages of Soybean Development**. 2nd. ed. Ames, Iowa: Iowa State University Press, 1977.

Governo do Estado do Rio Grande do Sul. **MAPA aceita proposta do RS de ampliar calendário de semeadura da soja**. 2024. Acesso em: 25 abr. 2025. Disponível em: <<https://www.agricultura.rs.gov.br/mapa-aceita-proposta-do-rs-de-ampliar-calendario-de-semeadura-da-soja>>.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann, 2011. 832 p.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

HYNDMAN, R. J.; KOHLER, A. B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, v. 22, n. 4, p. 679–688, 2006.

Instituto Brasileiro de Geografia e Estatística - IBGE. **Divisão regional do Brasil em regiões geográficas imediatas e regiões geográficas intermediárias: 2017**. Rio de Janeiro: IBGE, Coordenação de Geografia, 2017. 80 p. Coleção Ibgeana. Bibliografia nas p. [53]-58. ISBN 9788524044182. Disponível em: <<https://biblioteca.ibge.gov.br/visualizacao/livros/liv100600.pdf>>.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Produção de arroz no Rio Grande do Sul**. 2025. Acesso em: 30 abr. 2025. Disponível em: <<https://www.ibge.gov.br/explica/producao-agropecuaria/arroz/rs>>.

_____. **Produção de soja no Rio Grande do Sul**. 2025. Acesso em: 30 abr. 2025. Disponível em: <<https://www.ibge.gov.br/explica/producao-agropecuaria/arroz/rs>>.

LE, X.-H. et al. Application of long short-term memory (lstm) neural network for flood forecasting. **Water**, v. 11, n. 7, p. 1387, 2019.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. Statistical and machine learning forecasting methods: Concerns and ways forward. **PLOS ONE**, v. 13, n. 3, p. e0194889, 2018.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. 414 p.

MONTEITH, J. L. Climate and the efficiency of crop production in Britain. **Philosophical Transactions of the Royal Society of London. B, Biological Sciences**, The Royal Society, v. 281, n. 980, p. 277–294, 1977.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 5. ed. [S.l.]: Wiley, 2012.

NASA Center for Climate Simulation. **Earth System Grid Federation (ESGF) Climate Data Services**. 2024. Acesso em: 28 abr. 2025. Disponível em: <<https://www.nccs.nasa.gov/services/climate-data-services/esgf>>.

Nutrição de Safras. **Estádios fenológicos: entenda a fenologia das plantas**. 2025. <https://nutricaoadesafras.com.br/estadios-fenologicos-fenologia>. Acesso em: 30 abr. 2025.

OLIVEIRA, I. et al. A scalable machine learning system for pre-season agriculture yield forecast. **IBM Research**, 2021. IBM Research.

OLIVEIRA, R. F.; CUNHA, G. R.; STRECK, N. A. Influência da umidade do solo na produtividade da soja em diferentes cenários climáticos no sul do Brasil. **Pesquisa Agropecuária Brasileira**, v. 50, n. 11, p. 1012–1020, 2015.

O'NEILL, B. C. et al. The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. **Global Environmental Change**, v. 42, p. 169–180, 2017.

PEREIRA, A. R.; ANGELOCCI, L. R.; SENTELHAS, P. C. **Agrometeorologia: fundamentos e aplicações práticas**. Guaíba: Editora Guaíba, 2002.

PHILIPP, G.; SONG, D.; CARBONELL, J. G. The exploding gradient problem demystified: Definition, prevalence, impact, origin, tradeoffs, and solutions. **arXiv preprint arXiv:1712.05577**, 2017.

RIAH, K. et al. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. **Global Environmental Change**, v. 42, p. 153–168, 2017.

Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação do Rio Grande do Sul. **Radiografia da Agropecuária Gaúcha 2024**. Porto Alegre: Governo do Estado do Rio Grande do Sul, 2024. Coordenação: Paulo Lipp João. Acesso em: 25 abr. 2025. Disponível em: <<https://www.agricultura.rs.gov.br/upload/arquivos/202408/26113434-rag-2024-22-08-24-final-cap-a-atualizada.pdf>>.

SENTELHAS, P. C.; BATTISTI, R.; CÂMARA, G. M. S. Agrometeorologia aplicada ao manejo da irrigação e ao zoneamento agrícola de risco climático. **Revista Brasileira de Engenharia Agrícola e Ambiental**, UFPB, v. 19, n. S1, p. 1–10, 2015.

TECHNOLOGIES, C. **Random Forest vs Decision Tree: Difference Between Random Forest and Decision Tree**. 2024. Acesso em: 30 abr. 2025. Disponível em: <<https://codalien.com/blog/random-forest-vs-decision-tree-guide/>>.

VASCONCELOS, E. S.; SILVA, L. A. d. Análise estatística e modelos de machine learning na produção agrícola brasileira: Tendências temporais e eficiência produtiva ao longo de quatro décadas (1980-2019). **Instituto Federal Goiano, Goiânia, Goiás Brasil**, 2021. Doutor em Engenharia Elétrica, Instituto Federal Goiano; Doutor em Engenharia Elétrica, Universidade de Uberaba (UNIUBE), Uberaba, Minas Gerais, Brasil.

World Climate Research Programme. **World Climate Research Programme**. 2024. Acesso em: 25 abr. 2025. Disponível em: <<https://www.wcrp-climate.org/>>.

YAN, S. **Understanding LSTM and its diagrams**. 2016. Acesso em: 30 abr. 2025. Disponível em: <<https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714>>.

ZHOU, Z.-H. **Machine learning**. [S.l.]: Springer nature, 2021. 542 p.