MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m12e/2025/05.16.01.02-TDI

ATRIBUTOS NATURAIS QUE MODULAM A OCORRÊNCIA DE FORMAÇÕES VEGETAIS NO BIOMA CERRADO: UMA ABORDAGEM A PARTIR DO APRENDIZADO DE MÁQUINA EM BASES DE DADOS DE INFORMAÇÕES AMBIENTAIS

Alda Monteiro Barbosa

Tese de Doutorado do Curso de Pós-Graduação em Ciência do Sistema Terrestre, orientada pelo Dr. Jean Pierre Henry Balbaud Ometto, aprovada em 07 de abril de 2025.

URL do documento original:
<http://urlib.net/8JMKD2USNRW34T/4DG264P>

INPE
São José dos Campos
2025

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELECTUAL DO INPE - CEPPII (PORTARIA Nº 176/2018/SEI-INPE):

Presidente:

Dr. Thales Sehn Korting - Coordenação-Geral de Ciências da Terra (CGCT)

Membros:

Dr. Antonio Fernando Bertachini de Almeida Prado - Conselho de Pós-Graduação (CPG)

Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia, Tecnologia e Ciência Espaciais (CGCE)

Dr. Heyder Hey - Coordenação-Geral de Infraestrutura e Pesquisas Aplicadas (CGIP)

Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon

Clayton Martins Pereira - Divisão de Biblioteca (DIBIB)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)

André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)

EDITORAÇÃO ELETRÔNICA:

Ivone Martins - Divisão de Biblioteca (DIBIB)

André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m12e/2025/05.16.01.02-TDI

ATRIBUTOS NATURAIS QUE MODULAM A OCORRÊNCIA DE FORMAÇÕES VEGETAIS NO BIOMA CERRADO: UMA ABORDAGEM A PARTIR DO APRENDIZADO DE MÁQUINA EM BASES DE DADOS DE INFORMAÇÕES AMBIENTAIS

Alda Monteiro Barbosa

Tese de Doutorado do Curso de Pós-Graduação em Ciência do Sistema Terrestre, orientada pelo Dr. Jean Pierre Henry Balbaud Ometto, aprovada em 07 de abril de 2025.

URL do documento original:
<http://urlib.net/8JMKD2USNRW34T/4DG264P>

INPE
São José dos Campos
2025

Dados Internacionais de Catalogação na Publicação (CIP)

Barbosa, Alda Monteiro.

B234a Atributos naturais que modulam a ocorrência de formações vegetais no Bioma Cerrado: uma abordagem a partir do aprendizado de máquina em bases de dados de informações ambientais / Alda Monteiro Barbosa. – São José dos Campos : INPE, 2025.

xxvi + 133 p. ; (sid.inpe.br/mtc-m12e/2025/05.16.01.02-TDI)

Tese (Doutorado em Ciência do Sistema Terrestre) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2025.

Orientador : Dr. Jean Pierre Henry Balbaud Ometto.

1. Cerrado. 2. Aprendizado de máquina. 3. Grade estatística do IBGE. 4. Modelos híbridos. 5. Mapa de potencialidade natural I.Título.

CDU 502.5:004.032.2(213.54)



Esta obra foi licenciada sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Serviço de Pós-Graduação - SEPGR

Aluno (a): Alda Monteiro Barbosa

Novo Título: "Atributos naturais que modulam a ocorrência de formações vegetais no Bioma Cerrado: uma abordagem a partir do aprendizado de máquina em Bases de Dados de Informações Ambientais"

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título Doutor (a) em Ciência do Sistema Terrestre.

(assinado eletronicamente)
Dra. Silvana Amaral Kampel - Presidente - INPE
(x) Aprovado ( ) Reprovado

(assinado eletronicamente)
Dr. Jean Pierre Henry Balbaud Ometto - Orientador - INPE
(x) Aprovado ( ) Reprovado

(assinado eletronicamente)
Dr. Manoel Ferreira Cardoso - Membro Interno - INPE/DIOTG
(x) Aprovado ( ) Reprovado

(assinado eletronicamente)
Dr. Giovanni de Araujo Boggione - Membro Externo - IFG/Coord. Geomática
(x) Aprovado ( ) Reprovado

(assinado eletronicamente)
Dra. Maira Barberi-Membro Externo - UCG/Depto. Paleoecologia
(x) Aprovado ( ) Reprovado

gov.br Documento assinado digitalmente
MAIRA BARBERI
Data: 22/09/2025 12:04:44-0300
Verifique em https://validar.iti.gov.br

Este trabalho foi aprovado por:

( ) maioria simples
(x) unanimidade

São José dos Campos, 07 de abril de 2025

sei! assinatura eletrônica

Documento assinado eletronicamente por **Giovanni de Araujo Boggione (E)**, **Usuário Externo**, em 09/09/2025, às 12:15 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jean Pierre Henry Balbaud Ometto**, **Chefe da Divisão de Projeto Estratégico 3**, em 09/09/2025, às 13:15 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Silvana Amaral Kampel**, **Pesquisadora**, em 18/09/2025, às 10:27 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Manoel Ferreira Cardoso**, **Pesquisador**, em 18/09/2025, às 11:30 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.mcti.gov.br/verifica.html, informando o código verificador **13126161** e o código CRC **EB1A3195**.

Referência: Processo nº 01340.001440/2025-52 SEI nº 13126161

“Quero ser como o relógio de sol: só registrar os momentos de luz”.

A autora.

Outono de 2025.

A meus pais:
Raimundo Expedito Mota Barbosa (*in memoriam*) e
Maria Adelaide Monteiro Barbosa (*in memoriam*),

à minha querida irmã **Adélia Monteiro Barbosa** (*in memoriam*),
à minha pequena e doce sobrinha **Emanuela Barbosa Perdigão** (*in memoriam*)
e
à minha querida sogra mãe-amiga **Maria Vanda Ferreira** (*in memoriam*),

meus exemplos de amor e alegria pela vida, DEDICO.

AGRADECIMENTOS

Agradeço a Deus por ter me concedido vida, saúde e disposição para concluir o trabalho. Aos meus antepassados, sagrados espíritos, sempre me orientando e protegendo.

Agradeço ao meu orientador, Dr. Jean Pierre Ometto, pela compreensão, confiança e paciência.

Agradeço ao apoio financeiro desta pesquisa ao Instituto Brasileiro de Geografia e Estatística (IBGE) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, no Brasil (CAPES).

Agradeço aos amigos do IBGE, em especial à Drª Luciana Temponi, que me confiou a tarefa de colocar em desenvolvimento a ideia de onde surgiu esta pesquisa.

Agradeço ao meu chefe Dr. Pedro Edson e ex-chefe, Dr. Mario Ivan, pelo apoio, sugestões e compreensão.

Agradeço à equipe da RECOR-IBGE, especialmente nas pessoas de Frederico Takahashi e Alessandra Gouveia, pela gentileza de fornecerem todo material digital de mapas para a pesquisa.

Agradeço aos pesquisadores da banca examinadora: Drª Silvana Amaral, Dr. Dalton Valeriano (*in memoriam*), Dr. Giovanni Boggione, Drª Maira Barberi, Dr. Manoel Cardoso e Dr. Matheus Lima, pelo aceite do convite e pela compreensão.

Agradeço aos meus amigos ibgeanos Joana D'Arc, Luís Dambrós, Otto Neves, pela amizade, fotos de campo e contribuições diversas.

Agradeço à Ângela Harada, grande apoio na Pós-Graduação do curso de Ciência do Sistema Terrestre, em todos os momentos.

Agradeço à equipe da Biblioteca do INPE, que não mediu esforços para me fornecer todo material bibliográfico necessário à pesquisa.

Agradeço a(os) amigo(a)s inpeanos de doutorado, professores e pesquisadores da DIIAV e DIOTG, pela convivência e amizade. Em especial, ao Dr. Thales Körting, por algumas orientações sobre o Aprendizado de Máquina.

Agradeço aos familiares e inúmeros amigos que estiveram incentivando e torcendo o tempo todo por mim, nos momentos de dificuldade que enfrentei.

Agradecimento imenso ao meu filho Gustavo, companheiro inseparável de todas as horas, que me incentivou o tempo todo. Filho maravilhoso, motivo maior de eu ter enfrentado este imenso desafio! Muito obrigada!

RESUMO

O cerrado, a savana brasileira, é um *hotspot* de biodiversidade vital para o equilíbrio ambiental. Atualmente, é um dos biomas mais ameaçados do Brasil, exigindo ações urgentes de conservação e restauração. A pesquisa investigou a possibilidade de algoritmos "aprenderem" quais padrões associativos entre alguns atributos físicos que ocorrem na natureza estão relacionados à formação da vegetação de cerrado. Padrões associativos que ocorrem naturalmente, no decorrer do tempo, não são tão perceptíveis à análise convencional de mapeamento para um usuário comum. A fim de compreender os achados científicos sobre fatores ambientais que possam modular a ocorrência do cerrado, a revisão teórica desta pesquisa abordou a evolução do cerrado a partir do Último Máximo Glacial. O delineamento experimental utilizou-se de técnicas de mineração de dados e de aprendizado de máquina, a partir de atributos ambientais integrados na grade estatística de informações ambientais do IBGE (1 km x 1 km), para identificar padrões associativos entre variáveis ambientais na ocorrência de formações de savanas. Dados de clima, geologia, geomorfologia, solos, água disponível no solo, declividade e as tipologias de savanas do Bioma Cerrado foram provenientes dos levantamentos de recursos naturais disponíveis nos repositórios de informações ambientais oficiais do IBGE, apoiados com dados da EMBRAPA e CPRM. Visando regiões de vegetação mais preservada, mas com levantamentos realizados e integrados por especialistas temáticos, a pesquisa utilizou, como área-piloto, uma área *core* do cerrado: a Reserva Ecológica (RECOR) do IBGE. Inicialmente, explorou-se a aplicação de regras de associação (*Association Rules*) para identificar relações frequentes entre os atributos ambientais. Entretanto, percebeu-se que os resultados das regras de associação demonstravam apenas coocorrência entre os atributos, sem estabelecer uma relação causal ou preditiva. Diante desta limitação, a pesquisa evoluiu para o uso de Árvores de Decisão (J48) e *Random Forest* (RF) como algoritmos classificadores, buscando aumentar a capacidade preditiva do modelo e expandir para análise das classes de savana do Bioma Cerrado (Savana Florestada, Savanas Parque, Savanas Arborizadas e Savanas Gramíneo-Lenhosas). Apesar do bom desempenho dos algoritmos J48 e RF (79,16% e 83,65% de acurácia global, respectivamente), a classe minoritária das tipologias de savana, a Savana Parque, não foi suficientemente discriminada. A integração com Redes Bayesianas (BN) permitiu a incorporação de inferência probabilística, resultando em um modelo híbrido capaz de melhorar a acurácia e a compreensão das relações entre os atributos ambientais e os tipos de vegetação do cerrado. Como resultado, o classificador híbrido final (RF + BN) atingiu uma taxa de acerto de 72,26%, porém, discriminando melhor a Savana Parque no teste de sensibilidade (maior Recall), demonstrando a viabilidade da abordagem proposta. Os atributos ambientais que mais responderam para a ocorrência do cerrado foram os relacionados ao relevo e clima. Futuros desdobramentos desta pesquisa podem apoiar a elaboração de um mapa de potencialidade natural do cerrado e de estudos para a recomposição do cerrado.

Palavras-chave: Cerrado. Aprendizado de Máquina. Grade Estatística do IBGE. Modelos Híbridos. Mapa de potencialidade natural.

NATURAL ATTRIBUTES DRIVING VEGETATION FORMATION OCCURRENCE IN THE CERRADO BIOME: A MACHINE LEARNING-BASED ANALYSIS OF ENVIRONMENTAL DATABASES

ABSTRACT

The Cerrado, the Brazilian savanna, is a biodiversity hotspot crucial for environmental balance. Currently, it is one of Brazil's most threatened biomes, demanding urgent conservation and restoration efforts. This research investigates the potential of algorithms to identify associative patterns among natural attributes, particularly those related to vegetation formation. Many of these patterns are not easily discernible through conventional mapping analysis. The study focuses on applying data mining and machine learning techniques to integrated environmental attributes from IBGE's statistical grid of environmental information (1 km x 1 km). Predictive algorithms were employed to detect associative patterns between environmental variables and Cerrado vegetation classes. The data originated from official IBGE repositories on natural resources, complemented by EMBRAPA and CPRM datasets. Climate, geology, geomorphology, soil types, Available Water Capacity (AWC), slope, and savanna vegetation typologies from the Cerrado biome were used in the experimental design. A theoretical review on the Cerrado's evolution since the Last Glacial Maximum was conducted to explore environmental factors that may influence its occurrence. To identify well-preserved vegetation elements with integrated thematic surveys, the IBGE Ecological Reserve (RECOR), a core Cerrado area, was selected as a pilot. Initially, Association Rules were applied to identify frequent relationships among environmental attributes. However, this method only demonstrated cooccurrence without establishing causal or predictive relationships. To address this limitation, Decision Trees (J48) and Random Forest were utilized as classifiers to enhance predictive capacity and analyze the Cerrado savanna classes: Forested Savanna, Park Savannas, Woodland Savannas, and Grass-Wooded Savannas. Despite strong performance by J48 and Random Forest algorithms (79.16% and 83.65% overall accuracy, respectively), some savanna types were not sufficiently differentiated. Integrating Bayesian Network (BN) further incorporated probabilistic inference, resulting in a hybrid model that improved accuracy and enhanced the understanding of relationships between environmental attributes and Cerrado vegetation types. The final hybrid classifier (RF + BN) achieved an accuracy rate of approximately 72.26%, providing better discrimination (higher recall or sensitivity) of Park Savannas, a minority class among the savanna typologies, demonstrating the feasibility of the proposed approach. Topography and climate-related factors were the most influential environmental drivers of the Cerrado's distribution. Future developments of this research may support the creation of a natural potential map of the Cerrado and studies for its preservation and restoration.

Keywords: Cerrado. Machine learning. IBGE's Statistical grid. Hybrid model. Natural potential map.

LISTA DE FIGURAS

Pág.

Figura 2.1- Forçantes Climáticas..... 8
Figura 2.2 – Precessão dos equinócios. 9
Figura 2.3 – Mudanças na obliquidade da Terra. 10
Figura 2.4 – Excentricidade da órbita terrestre..... 11
Figura 2.5 - Mapa de Biomas e Sistema Costeiro-Marinho do Brasil. 18
Figura 2.6 - Domínios naturais da América do Sul. 24
Figura 2.7 - Diferença climática entre o UMG (20 mil anos AP) e a atualidade, considerando a temperatura e a precipitação. 26
Figura 3.1 - Esquema do processo KDD (Knowledge Discovery in Dataset). .. 33
Figura 3.2 – Interface do WEKA, onde aparece a ave da Nova Zelândia que dá origem ao nome do aplicativo..... 35
Figura 4.1 – Portal do BDiA do IBGE. 40
Figura 4.2 – Substrato Geológico do Bioma Cerrado. 42
Figura 4.3 – Compartimentação de relevo do Bioma Cerrado. 43
Figura 4.4 – Solos do Bioma Cerrado. 44
Figura 4.5 – Subgrupo de Formação de vegetação do Bioma Cerrado. 45
Figura 4.6 - Blocos-Diagramas das fisionomias ecológicas da Savana, segundo o IBGE..... 48
Figura 4.7 - Zonas climáticas para o Cerrado, segundo Nimer (1979). 49
Figura 4.8 – Subzonas climáticas para o Bioma Cerrado, segundo Nimer (1979). 50
Figura 4.9 - Visão da Grade Estatística de Dados Ambientais, do IBGE, com o recorte do Bioma Cerrado. 52
Figura 4.10 – Padrão de cores e intervalos das classes de Água Disponível no solo..... 54
Figura 4.11 – Classes de Declividade em percentagem e suas cores. 56
Figura 4.12 – Cartas de Declividade por articulação ao Milionésimo. 57
Figura 5.1 – Ortocarta de localização geográfica da RECOR. 60
Figura 5.2 – Integração de temas da RECOR para o SIG. 62
Figura 5.3 – Esquema dos processamentos realizados na pesquisa..... 72

Figura 6.1 – Visão da Árvore de Decisão (algoritmo J48) para os dados da RECOR. 80
Figura 6.2 – Análise MCA para os tipos de Savana do Bioma Cerrado. 82
Figura 6.3 – Espacialização do algoritmo Random Forest para o Bioma Cerrado (amostragem de 10%). 89
Figura 6.4 – Probabilidades consolidadas, por variável ambiental para ocorrência de Savanas no Bioma Cerrado (valores acima de 50%). 94

LISTA DE TABELAS

Pág.

Tabela 2.1 - Glaciações do Quaternário. 13
Tabela 2.2 – Eventos e períodos de alterações climáticas abruptas ocorridos nos últimos 21.000 anos AP. 16
Tabela 4.1 – Dados ambientais, instituição do repositório, escala/resolução espacial e endereço da url. 41
Tabela 6.1 – Síntese das Regras de Associação para a RECOR. 78
Tabela 6.2 – Síntese dos principais padrões associativos dos atributos pelas Regras de Associação. 84
Tabela 6.3 – Resultado das métricas de desempenho do classificador J48. ... 86
Tabela 6.4 – Desempenho do algoritmo classificador J48 para as classes do Bioma Cerrado. 87
Tabela 6.5 – Desempenho por classe de Bioma Cerrado para o algoritmo RF (50% de amostragem estratificada). 90
Tabela 6.6 - Desempenho por classe do classificador BN para amostragem de 50% do Bioma Cerrado. 91
Tabela 6.7 – Comparação entre os desempenhos individuais dos classificadores Random Forest e Bayes Net para os dados do Bioma Cerrado. 95
Tabela 6.8 - Precisão Detalhada Por Classe do Modelo Hibrido RF + BN. 96

LISTA DE SIGLAS E ABREVIATURAS

AP	Antes do Presente
AMOC	*Atlantic Meridional Overturning Circulation*
AR	*Association Rules*
A-ZCIT	Zona de Convergência Intertropical do Atlântico
BA	*Bølling-Alerød*
BDiA	Banco de Dados de Informações Ambientais
BN	*Bayes Net*
C3	3-*Carbon compound*
C4	4-*Carbon compound*
CO_2	Dióxido de Carbono
CPRM	Companhia de Pesquisa de Recursos Minerais
D-O	*Dansgaard-Oeschger*
DNA	*Deoxyribonucleic Acid*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
H1	*Heinrich Stadial 1*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
INDE	Infraestrutura Nacional de Dados Espaciais
INPE	Instituto Nacional de Pesquisas Espaciais
KDD	*Knowledge Discovery in Database*
LGM	*Last Glacial Maximum*
MCA	*Multiple Correspondence Analysis*
PEC	Padrão de Exatidão Cartográfica
RECOR	Reserva Ecológica do IBGE
RF	Random Forest
RIDE-DF	Região Integrada de Desenvolvimento do Distrito Federal e Entorno
SIG	Sistema de Informações Geográficas
SMAS	Sistema de Monções da América do Sul
SMOTE	*Synthetic Minority Over-sampling Technique*
SRTM	*Shuttle Radar Topography Mission*
TraCE-21k	*Transient Climate Evolution over the last 21.000 years*

UMG	Último Máximo Glacial
UTB	Unidade Territorial Básica
YD	*Younger Dryas*
ZCIT	Zona de Convergência Intertropical
ZEE	Zoneamento Ecológico-Econômico

LISTA DE SÍMBOLOS

°C Graus Celsius

± Aproximadamente, mais ou menos.

SUMÁRIO

Pág.

1 INTRODUÇÃO 1
1.1 Objetivos 3
1.1.1 Objetivo geral 3
1.1.2 Objetivos específicos 4
2 EVOLUÇÃO E DISTRIBUIÇÃO DO CERRADO EM UM CONTEXTO BIOGEOGRÁFICO E PALEOCLIMÁTICO: UMA REVISÃO DESDE O ÚLTIMO MÁXIMO GLACIAL (UMG) 5
2.1 Clima do passado e as glaciações 5
2.1.1 O sistema climático e suas forçantes 6
2.1.1.1 Mudanças na órbita da Terra 8
2.1.1.1.1 Movimentos de precessão 9
2.1.1.1.2 Mudanças na obliquidade da Terra 10
2.1.1.1.3 Mudanças na excentricidade da órbita terrestre 10
2.1.1.2 Mudanças na atividade solar 11
2.1.1.3 Movimentação das placas tectônicas 12
2.1.2 O clima e os *feedbacks* 12
2.1.3 O clima do Quaternário 12
2.1.4 Os eventos de mudanças climáticas abruptas do Quaternário 15
2.2 O Cerrado e suas diversas origens 16
2.2.1 O debate acadêmico sobre as origens do cerrado 19
2.2.2 Aspectos das origens do cerrado e o fogo 20
2.2.3 Teria o cerrado uma origem mais antiga? 21
2.3 Processos paleoclimáticos e a dinâmica biogeográfica do Brasil, no Último Máximo Glacial 22
2.3.1 Último Máximo Glacial mais frio e seco ou mais frio e úmido? 22
2.3.1.1 Hipótese do UMG mais frio e seco 23
2.3.1.2 Hipótese do UMG mais frio e úmido 25
2.3.2 O registro de pólen no paleoambiente e a evolução do cerrado 27
2.4 Considerações da revisão para os objetivos da pesquisa. 29
3 FUNDAMENTAÇÃO TEÓRICA DO DELINEAMENTO EXPERIMENTAL . 32

3.1 Fundamentação teórica de machine learning, mineração de dados e suas aplicações 32
3.2 O método KDD: descoberta de conhecimento em bases de dados 33
3.3 QGIS 34
3.4 WEKA 35
3.5 Regras de associação e suas métricas 36
3.6 Árvores de Decisão 37
3.7 Random Forest 37
3.7.1 Diferenças principais entre Árvore de Decisão e Random Forest 38
3.8 Redes bayesianas 38
3.8.1 Vantagens de redes bayesianas 38
3.9 Métricas de desempenho dos algoritmos classificadores 39
4 INFORMAÇÕES AMBIENTAIS EM REPOSITÓRIOS INSTITUCIONAIS SOBRE O CERRADO UTILIZADAS NA PESQUISA 40
4.1 Informações ambientais do BDiA 41
4.1.1 Geologia 41
4.1.2 Geomorfologia 42
4.1.3 Solos 43
4.1.4 Vegetação 45
4.1.4.1 As unidades de Savana no BDiA 46
4.1.4.1.1 A Savana florestada 47
4.1.4.1.2 Savana arborizada 47
4.1.4.1.3 Savana parque 47
4.1.4.1.4 Savana gramíneo-lenhosa (Campo-Limpo-de-Cerrado) 48
4.2 Informações ambientais de Clima - IBGE 48
4.3 Grade estatística de informações ambientais - IBGE 51
4.4 Informações ambientais da EMBRAPA 52
4.4.1 Água disponível no solo 52
4.4.2 Definição das classes de Água Disponível (AD) no solo 53
4.5 Informações ambientais do CPRM/SGB 55
4.5.1 Classificação de declividade da EMBRAPA 55
4.5.2 Dados de declividade da CPRM 56
5 MATERIAIS E MÉTODOS 58
5.1 Descrição da área da RECOR 59

5.2 A etapa experimental com os dados da RECOR 60
5.3 A etapa experimental com os dados do Bioma Cerrado 64
5.4 As classes de variáveis preditoras em cada experimento 67
5.4.1 Variáveis da RECOR 68
5.4.2 Variáveis do Bioma Cerrado 70
5.5 Síntese dos processamentos 71
6 RESULTADOS E DISCUSSÃO 73
6.1 Resultados da RECOR 73
6.1.1 Regras de associação com os dados da RECOR 73
6.1.1.1 Análise dos resultados da RECOR para as regras de associação, por cada classe de vegetação 74
6.1.1.2 Tabela síntese das regras de associação para a RECOR 78
6.1.2 J48 para os dados da RECOR 79
6.1.3 RF para os dados da RECOR 80
6.2 Resultados do Bioma Cerrado 81
6.2.1 A análise MCA para as savanas do Bioma Cerrado 81
6.2.2 Regras de associação para os dados do Bioma Cerrado 83
6.2.2.1 Tabela síntese das regras de associação para o Bioma Cerrado 83
6.2.2.2 Padrões gerais nas associações de Savanas do Cerrado 86
6.2.3 J48 para os dados do Bioma Cerrado 86
6.2.4 RF para os dados do Bioma Cerrado 88
6.2.5 BayesNet (BN) para os dados do Bioma Cerrado 90
6.2.6 Classificador híbrido (RF + BN) para o Bioma Cerrado 94
7 CONSIDERAÇÕES FINAIS 98
REFERÊNCIAS BIBLIOGRÁFICAS 101
GLOSSÁRIO DE TERMOS DE MACHINE LEARNING UTILIZADOS NA PESQUISA 110
APÊNDICE A – VARIÁVEIS-ALVO E PREDITORAS DA RECOR 113
APÊNDICE B – VARIÁVEIS-ALVO E PREDITORAS DO BIOMA CERRADO 114
APÊNDICE C – EXEMPLOS DE VISUALIZAÇÂO DAS REGRAS DE ASSOCIAÇÃO NO WEKA 115
APÊNDICE D – GRAFO DA REDE BAYESIANA COM TODAS AS VARIÁVEIS AMBIENTAIS 116

APÊNDICE E – EXEMPLOS DE TABELAS DE DISTRIBUIÇÃO DE PROBABILIDADE (TDP) GERADAS PELO CLASSIFICADOR BAYES NET 117
APÊNDICE F - EXEMPLO DE REGRAS DE ASSOCIAÇÃO GERADAS NO WEKA 118
APÊNDICE G – CÓDIGOS PYTHON UTILIZADOS NA PESQUISA 122
CLASSIFICADOR DE REDES BAYESIANAS 122
APÊNDICE H – CÓDIGO PYTHON DAS REDES BAYESIANAS, TABELAS E GRÁFICOS DE PROBABILIDADE 125
APÊNDICE I – CÓDIGO PYTHON PARA ANÁLISE MCA 131
ANEXO A – CARTA CRONOESTRATIGRÁFICA INTERNACIONAL 133

1 INTRODUÇÃO

O Cerrado é uma das regiões de maior biodiversidade do mundo. Ao lado da Mata Atlântica, é considerado um dos *hotspots* mundiais de biodiversidade. O cerrado também é considerado como a “caixa d’água” dos principais mananciais brasileiros. Afinal, oito das dozes bacias hidrográficas brasileiras possuem nascentes neste bioma. É nele que são abastecidos os aquíferos Bambuí, Urucuia e Guarani. O bioma Pantanal é totalmente dependente da água do Cerrado e grande parte da energia consumida no Brasil é gerada com as águas do Cerrado (REDE CERRADO, 2021).

O homem ocupa o Cerrado há mais de 11 mil anos antes do presente[1] (AP) (BARBOSA, 1995; 2008). Os povos originários do Cerrado são constituídos por uma variedade considerável de grupos com diferentes referências culturais. Atualmente, a área deste bioma abriga 95 territórios indígenas, 44 territórios quilombolas e 13 tipos de comunidades tradicionais não indígenas, que fornecem um testemunho vivo da rica tradição de convivência humana com a natureza.

A presente pesquisa visa contribuir para o entendimento da formação e ocorrência do cerrado. Apesar das pesquisas sobre o tema terem avançado em muitos questionamentos, a evolução e distribuição do cerrado ainda apresentam muitas lacunas que precisam ser melhor estudadas, visto que ainda são motivos de muita controvérsia científica.

De um modo geral, a discussão sobre a evolução da cobertura vegetal brasileira como resposta às mudanças climáticas tem ganhado cada vez mais contribuições de vários campos científicos. Merecem destaque os diversos ramos das paleociências, tais como: a Paleoclimatologia, a Paleobotânica (através da Palinologia), a Paleontologia e Arqueologia, ou mesmo de outras

[1] AP – Abreviatura para “Antes do Presente", onde “presente”, em geocronologia, é convencionalmente definido como o ano 1950 d.C. (para evitar confusão com datas baseadas em atividades nucleares modernas).

ciências afins, como a Geologia, Geomorfologia, Pedologia e a Biogeografia. Mais recentemente, uma nova área da Biologia Molecular tem contribuído nas pesquisas sobre a origem e dispersão de várias espécies, através da análise de sequenciamento de DNA.

A problematização que foi proposta nesta pesquisa esteve focada em uma abordagem de atributos naturais preditivos para ocorrência do cerrado. Através de padrões associativos que envolvem o clima, a geologia, a geomorfologia, solos e o tempo na formação de tipologias vegetais, a linha investigativa da pesquisa se baseou na possibilidade de algoritmos perceberem padrões associativos entres os atributos que ocorrem na natureza, em especial aqueles que estão relacionados à formação da vegetação. Muitos padrões associativos que ocorrem naturalmente não são tão perceptíveis à análise convencional de mapeamento para um usuário.

A percepção de padrões ecológicos não é algo tão trivial de ser apreendido pela enorme quantidade de processos envolvidos, no tempo e no espaço, para a formação de uma cobertura vegetal.

A partir da segunda metade do século passado, cada vez mais as ciências ambientais têm recebido uma enorme contribuição da área da ciência da computação para as análises relacionadas ao levantamento de um grande volume de dados. Neste aspecto, o advento de técnicas que envolvem o sensoriamento remoto, os Sistemas de Informações Geográficas, a estatística multivariada, a mineração de dados e, mais recentemente, diversos segmentos da inteligência artificial - como as redes neurais, o aprendizado de máquina, o aprendizado profundo - têm impulsionado cada vez mais o leque de conhecimento das ciências que estudam o meio ambiente.

Deste volume muito grande de informações e de novas técnicas emergentes, nasce a necessidade de muitas instituições, públicas ou não, que trabalham na área ambiental, saber utilizá-las da melhor forma possível. Daí nasceu a problematização desta pesquisa.

a) Seria possível determinar critérios de padrões associativos que ocorrem na natureza entre diversos atributos ambientais na formação de tipos

vegetacionais do cerrado, utilizando de técnicas de mineração de dados (*Data Mining*), aprendizado de máquina (*Machine Learning)* em uma análise temporal conhecida?

b) Se sim, como validar??

c) Seria possível definir padrões na natureza das condições para a ocorrência específica das formações de Savanas no Bioma Cerrado a partir de uma análise de uma série climática conhecida?

As hipóteses levantadas foram:

a) Sim, as atuais técnicas de mineração de dados e aprendizado de máquina tornam possível criar um modelo preditivo a partir da análise probabilística de padrões associativos que se apresentem em uma grande quantidade de amostras em uma classe temática (alvo) com critérios de regras apoiados pela literatura e por especialistas temáticos.

b) Uma vez que, teoricamente, exista uma interdependência entre os atributos ambientais como clima, geologia, geomorfologia, solos e tempo, é possível validar um classificador que estabeleça regras de associações confiáveis entre os atributos (*inputs*) fornecidos ao modelo, em um período de tempo que se conheça a série histórica do clima.

c) Uma vez tendo como entrada (*inputs*) do modelo um grande volume de tipologias de formações vegetais de savanas conhecidas do Bioma Cerrado, o modelo classificador aprenderá a discriminar as diferentes formações de savanas.

1.1 Objetivos

1.1.1 Objetivo geral

a) Propor um modelo classificador, baseado em padrões associativos, que possa identificar áreas potenciais para a ocorrência de tipos de vegetação do cerrado a partir de um conjunto de atributos ambientais fornecidos.

1.1.2 Objetivos específicos

a) Detectar padrões e estabelecer regras de associação entre variáveis ambientais, através de um algoritmo de mineração de dados, em uma área de cerrado, com condições ambientais controladas e/ou conhecidas.

b) Validar um modelo preditivo para a ocorrência de formações vegetais de savanas, a partir de variáveis ambientais de diversas bases de dados ambientais levantadas em áreas de ocorrência do Cerrado.

c) Validar um modelo classificador para as savanas do Bioma Cerrado.

O documento, de uma forma geral, discorre sobre a pesquisa da seguinte forma: há uma revisão de literatura, no contexto paleoclimático e biogeográfico, sobre a evolução do cerrado, especialmente desde o Último Máximo Glacial. Em seguida, uma descrição da fundamentação teórica sobre os conceitos metodológicos utilizados e a descrição do experimento propriamente dito. Depois serão descritas as análises, discussão e, finalmente, as considerações finais. Para facilitar o entendimento cronogeológico de alguns termos, no Anexo A do trabalho, está apresentada a Carta Cronoestratigráfica Internacional vigente para o período da pesquisa. Também há um glossário de alguns termos mais específicos do universo do Aprendizado de Máquina (*Machine Learning*) e um Apêndice com alguns exemplos de códigos na linguagem de programação Python utilizados no estudo, além da visualização de alguns dos resultados gerados para a análise da pesquisa.

No decorrer do trabalho, o leitor encontrará mais detalhes de como foram abordados e estruturados os tópicos relativos a cada capítulo.

2 EVOLUÇÃO E DISTRIBUIÇÃO DO CERRADO EM UM CONTEXTO BIOGEOGRÁFICO E PALEOCLIMÁTICO: UMA REVISÃO DESDE O ÚLTIMO MÁXIMO GLACIAL (UMG)

Este capítulo diz respeito a uma revisão de literatura sobre a origem e evolução do cerrado desde o Último Máximo Glacial. O enfoque desta revisão é de cunho paleoclimático e biogeográfico, procurando discorrer sobre a dinâmica das variações climáticas ocorridas nos últimos 20-21 mil anos AP, no Brasil. Aspectos envolvendo debates teóricos sobre a dinâmica da evolução do cerrado, em especial, relacionados à teoria dos refúgios florestais no UMG, a influência do fogo, os registros paleoclimáticos e seus reflexos na evolução do cerrado também são contextualizados para os objetivos da pesquisa.

O Último Máximo Glacial - UMG (ou *Last Glacial Maximum* – LGM, em inglês) é convencionalmente definido, a partir dos registros do nível do mar, como o mais recente intervalo da história da Terra no qual as camadas de gelo alcançaram seus volumes máximos de integração (CLARK et al., 2009). É um período propício para estudar como os ecossistemas respondem às mudanças climáticas. Durante o UMG, houve uma expansão máxima das geleiras, as concentrações de CO_2 eram aproximadamente metade das atuais, as temperaturas eram até 5°C mais frias nas terras baixas tropicais e os regimes de precipitação eram diferentes dos atuais (STUTE et al., 1995). Com base em estudos geomorfológicos e análises de pólen, foram deduzidos climas frios e secos nos períodos glaciais e quentes e úmidos nos períodos interglaciais, entretanto novas pesquisas e modelos climáticos estão reformulando esses entendimentos (ARRUDA; SCHAEFER, 2020).

2.1 Clima do passado e as glaciações

O clima é um atributo ambiental de muita relevância para o estabelecimento das formas de vida, especialmente a vegetal. Como o clima está em constante mudança, é responsável por tornar o planeta Terra tão ativamente dinâmico (ARRUDA; SCHAEFER, 2020). A variabilidade do clima abrange todas as escalas de tempo - desde as mais curtas, com duração de poucos dias, até as mais longas, compostas por centenas de milhões de anos.

Compreender tal variabilidade tem um papel fundamental para que se possa antecipar a dinâmica do sistema climático no futuro. Para um bom entendimento da dinâmica do sistema climático atual são necessários estudos e observações da variabilidade climática natural, do balanço de energia, da compreensão da circulação geral da atmosfera e de padrões de teleconexões climáticos. As teleconexões climáticas compreendem uma visão geral da circulação atmosférica, onde forçantes locais agem para influenciar regiões remotas da Terra. As análises de teleconexões têm sido usadas principalmente para estudar as flutuações de grande escala planetária e de baixa frequência, que ocorrem na atmosfera.

A ciência que estuda o clima no passado é a Paleoclimatologia. A Paleoclimatologia é definida por Oliveira et al. (2015) como o estudo dos climas durante qualquer parte do passado geológico, sendo que trata especificamente dos estados do clima em épocas anteriores ao período de registros instrumentais modernos. Também Garcia et al. (2015) apresentam definição semelhante ao afirmarem que a Paleoclimatologia é a ciência que estuda o clima do passado e suas variações ao longo da história da Terra. Para os autores pode ser considerado paleoclima tudo o que ocorreu antes do ser humano conseguir obter informações sobre o tempo e o clima com termômetros e outros aparelhos específicos utilizados em medições meteorológicas.

2.1.1 O sistema climático e suas forçantes

O clima pode ser entendido como um sistema. O ar, os oceanos, o gelo, o solo, e a vegetação são os principais componentes desse sistema. Qualquer mudança em um desses componentes acarreta em alteração nas condições do clima (GARCIA et al., 2015). É importante conhecer o sistema climático atual para entender que as alterações no clima não ocorrem de maneira isolada. Mudanças na temperatura do oceano podem influenciar diretamente nas mudanças dos padrões de chuva no continente. O El Niño é um exemplo de fenômeno climático que depende diretamente das condições oceânicas.

A Terra sempre passou por ciclos naturais de aquecimento e resfriamento, ou seja, mudanças climáticas naturais sempre ocorreram em nosso planeta. Na

história evolutiva da Terra, quem sustenta o funcionamento do sistema climático e tem a capacidade de provocar mudanças naturais são as denominadas Forçantes Climáticas.

Para Garcia et al. (2015) podem ser consideradas três tipos de Forçantes Climáticas fundamentais: 1) mudanças na atividade solar; 2) mudança na órbita da Terra; e 3) movimentação das placas tectônicas. Sampaio et al. (2008) complementam ainda com os seguintes fatores que induzem as mudanças climáticas naturais: a quantidade de aerossóis naturais (provenientes de fontes minerais, incêndios florestais de origem natural e o sal marinho), as erupções vulcânicas e fenômenos climáticos que podem modificar o clima localmente, como os furacões, tempestades violentas e os fenômenos El Niño e La Niña.

Para Battarbee et al. (2008), essas forçantes climáticas podem ser classificadas em externas (relacionadas aos movimentos orbitais e solar) e internas (vulcanismo e circulação oceânica). Todos os processos radiativos dependem fortemente da composição da atmosfera (gases, aerossóis, poeira, nuvens), que fornecem outro nível de complexidade. Como esses componentes estão fortemente conectados a produtos químicos, térmicos e mudanças dinâmicas que ocorrem na atmosfera, em escalas de tempo muito diferentes, qualquer mudança nesses processos complexos pode forçar o clima a mudar.

Figura 2.1- Forçantes Climáticas.



Fonte: Adaptado de Ruddiman (2003).

A Figura 2.1 mostra, de forma esquemática, como as forçantes climáticas atuam em cada componente do sistema climático e estes, por sua vez, nas variações do clima. Três forçantes climáticas atuam sobre os componentes climáticos causando as variações do clima. São elas: mudanças na órbita da Terra, mudanças nas placas tectônicas e mudanças na radiação solar.

As forçantes aparecem como fatores que impulsionam as mudanças climáticas. Elas influenciam os componentes do sistema climático a interagirem entre si, provocando mudanças no clima.

2.1.1.1 Mudanças na órbita da Terra

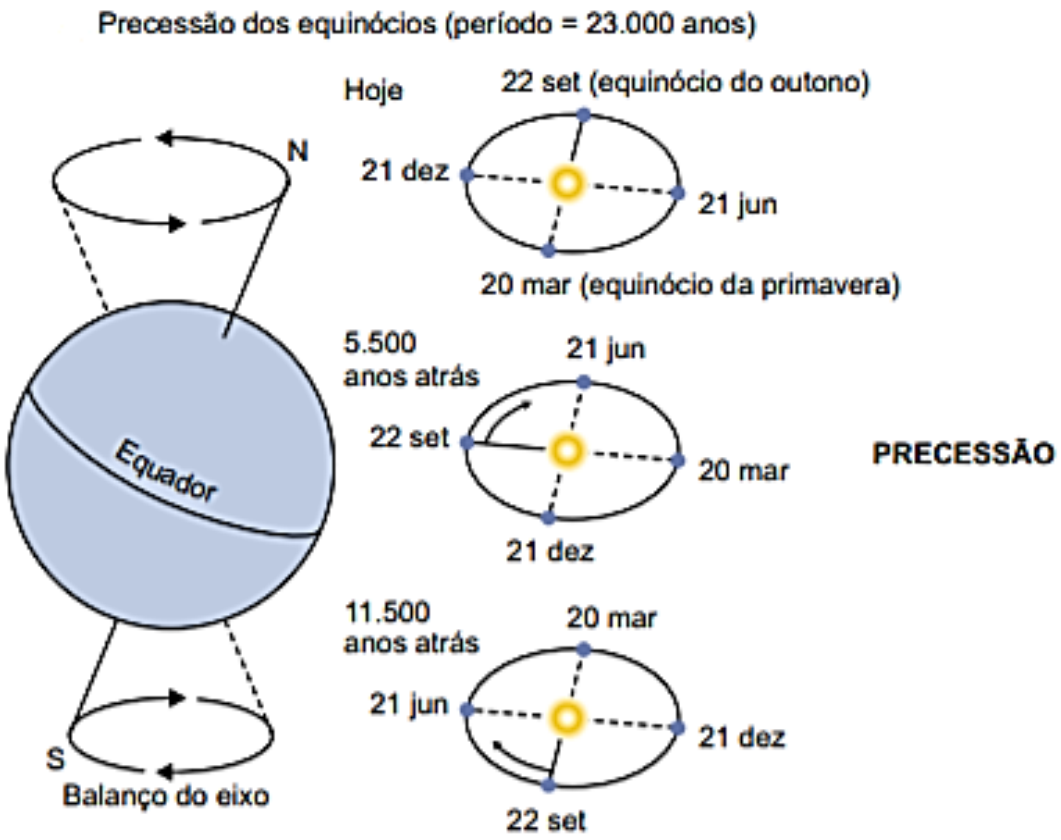
As alterações na órbita terrestre controlam a quantidade e distribuição de energia solar recebida pela Terra. A teoria matemática do clima baseada nas variações sazonais e latitudinais da radiação solar recebida pela Terra é conhecida como Teoria de Milankovitch**.** Recebeu este nome em homenagem ao cientista sérvio Mulatin Milankovitch (1879-1958), o qual teorizou estas variações cíclicas dos movimentos terrestres. Esta teoria se baseia nas variações cíclicas de três elementos, o que ocasiona variações na quantidade de energia solar que chega à Terra: a precessão, as mudanças de obliquidade e as variações na

excentricidade. Como estas alterações são cíclicas, também são conhecidas como Ciclos de Milankovitch. Segue uma breve descrição dessas variações cíclicas.

2.1.1.1.1 Movimentos de precessão

Os movimentos de precessão causam mudanças na orientação do eixo rotacional da Terra (Figura 2.2). Estas mudanças alteram as datas do periélio e do afélio e, portanto, aumentam o contraste sazonal em um Hemisfério e diminui em outro. O período médio do movimento de precessão é de 23.000 anos (MCGUFFIE; HENDERSON-SELLERS, 2005; SAMPAIO et al., 2008).

Figura 2.2 – Precessão dos equinócios.

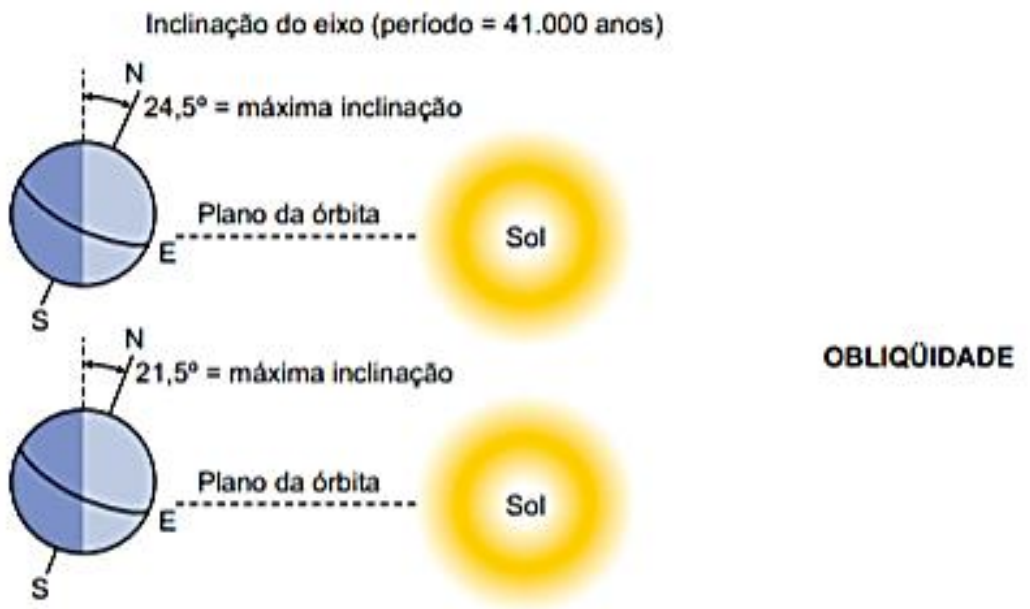


Fonte: Modificado de Sampaio et al. (2008).

2.1.1.1.2 Mudanças na obliquidade da Terra

As mudanças na obliquidade ocasionam mudanças na inclinação do eixo da Terra (Figura 2.3). Estas influenciam a magnitude da mudança sazonal, ou seja, quando a inclinação é maior, as estações são mais extremas (os invernos são mais frios e os verões mais quentes) e, quando a inclinação é menor, as estações são mais suaves em ambos os Hemisférios. Atualmente a inclinação é de 23,5°. O período médio das mudanças na obliquidade terrestre é de 41.000 anos, variando entre as inclinações de 21,5° e 24,5°. Quando os verões são mais frios, significa que há maior permanência de neve e gelo nas altas latitudes, o que contribui para um *feedback* positivo, ou seja, mais neve significa albedo maior e, portanto, maior resfriamento (CULVER; RAWSON, 2004; SAMPAIO et al., 2008).

Figura 2.3 – Mudanças na obliquidade da Terra.



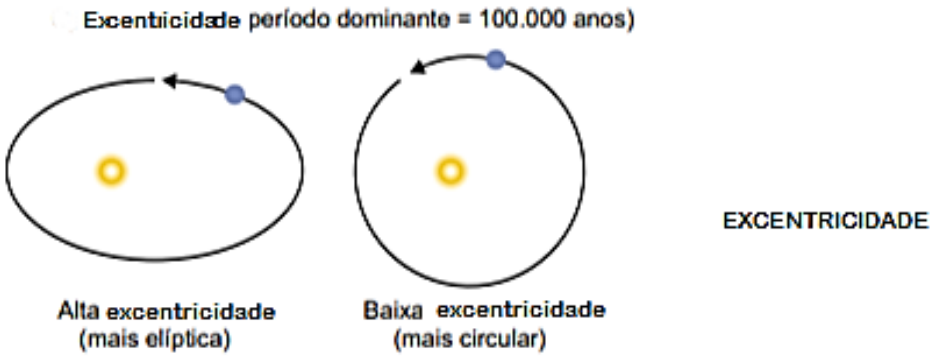
Fonte: Adaptado de Sampaio et al. (2008).

2.1.1.1.3 Mudanças na excentricidade da órbita terrestre

As variações na excentricidade estão associadas ao fato da órbita da Terra em relação ao Sol ser mais elíptica (alta excentricidade) ou mais circular (baixa excentricidade). Atualmente há uma diferença de 3% entre a maior aproximação

(periélio) e o afélio. Essa diferença na distância significa 6% da insolação entre janeiro e julho. Quando a órbita está mais elíptica, a diferença da insolação é da ordem de 20 a 30% entre janeiro e julho. O período das variações na excentricidade da órbita terrestre é de 90.000 a 100.000 anos (SAMPAIO et al., 2008). A Figura 2.4 apresenta a variação e o período de tempo de mudança na excentricidade da órbita da Terra.

Figura 2.4 – Excentricidade da órbita terrestre.



Fonte: Adaptado de Sampaio et al. (2008).

As variações na órbita da Terra, e as consequentes variações na quantidade de calor que é recebida nos diferentes locais, influenciaram a determinação das variações climáticas do passado. Com a variação da radiação solar por motivos astronômicos há variação no volume de gelo e, consequentemente, na temperatura global.

2.1.1.2 Mudanças na atividade solar

Mudanças na atividade solar afetam a quantidade de radiação solar recebida pela Terra, provocando aumento ou diminuição da temperatura global (MASSON-DELMOTTE et al., 2013; GARCIA et al., 2015). A atividade solar e, portanto, a radiação solar recebida pela Terra é caracterizada por ciclos regulares e recorrentes. Uma característica proeminente das mudanças na atividade solar está relacionada ao ciclo de manchas solares de 11 anos. Há indícios de que um período com baixo número de manchas solares está relacionado com condições climáticas mais frias (JOUSSAUME, 1996; MEYER; WAGNER, 2009).

2.1.1.3 Movimentação das placas tectônicas

O movimento das placas é capaz de originar oceanos e montanhas, que podem alterar o trajeto de correntes oceânicas, assim como perturbar o fluxo atmosférico oriundo de atividades vulcânicas, cujas atividades podem injetar na atmosfera gases do efeito estufa.

2.1.2 O clima e os *feedbacks*

A ação conjunta desses ritmos orbitais pode levar a variações cíclicas na intensidade e distribuição sazonal da radiação solar, influenciando a duração do período de degelo no verão, de modo que a neve depositada no inverno não se derreta completamente. Com o crescimento do manto de gelo, mais luz solar seria refletida (albedo), permitindo o aparecimento de grandes massas de gelo (ROCHA-CAMPOS; SANTOS, 2000). Esse efeito de crescimento do manto de gelo sendo incrementado pelo poder difusor do albedo permitindo o crescimento de mais massas de gelo é considerado um *feedback* positivo (ou de reforço) que amplifica o efeito da forçante climática. Salgado-Labouriau (1994) afirma que esses efeitos de *feedback* têm um equilíbrio delicado e, com uma pequena modificação, podem tender para um aquecimento ou resfriamento climático. Talvez seja sobre esse delicado equilíbrio que ocorram os mecanismos que desencadearam períodos estadiais ou interestadiais, cujas consequências são oscilações abruptas de temperatura.

Interessante ressaltar que vivemos em uma fase interglacial da idade glacial ocorrida na Era Cenozoica. Não se constitui uma novidade para a história geológica da Terra as variações climáticas drásticas que já ocorreram na superfície do planeta. De acordo com Rocha-Campos e Santos (2000), o registro geológico mostra evidências de pelo menos sete outras idades glaciais, bem documentadas, que se alternaram com períodos de aquecimento global.

2.1.3 O clima do Quaternário

Apesar da importância da compreensão dos climas do passado remoto (paleoclimas), as suas variações são muito lentas, quase imperceptíveis na escala humana. Por esse fato, muita relevância é dada ao período Quaternário,

ou seja, aos acontecimentos do último milhão de anos da Terra, quando surgem os primeiros ancestrais humanos que começam a produzir cultura e ter domínio do fogo (SANT'ANNA NETO; NERY, 2005; NEVES; RANGEL JÚNIOR; MURRIETA, 2015). Não foi possível, até o momento, encontrar evidências geológicas de todas as glaciações, pois muitas vezes uma glaciação se sobrepõe à outra, tornando o seu reconhecimento bastante difícil. E é em função desta dificuldade de registro e de conceituação que muitos autores divergem quanto ao número de glaciações ocorridas no passado geológico da Terra.

Salgado-Labouriau (1994) e Sant'Anna Neto e Nery (2005) resumem as cinco grandes glaciações conhecidas e detectadas no período Quaternário conforme demonstradas na Tabela 2.1, com destaque para a última glaciação, denominada de *Würm* para os europeus, ou *Wisconsin*, para os norte-americanos, ou ainda de *Würm-Wisconsiana*.

Tabela 2.1 - Glaciações do Quaternário.

Nome das Glaciações		**Período (em milhares de anos)**
Europa	**EUA**	
Donau	-	1600
Günz	Nebraskan	600-500
Mindel	Kansan	480-440
Riss	Illinoian	230-190
Würm	Wisconsin	115-12

Fonte: Adaptada de Sant'Ann e Nery (2005).

Convém ressaltar que a glaciação Würm-Wisconsiana começou muito antes do Último Máximo Glacial. Würm e Wisconsin são glaciações regionais que ocorreram em partes específicas do hemisfério Norte. Würm refere-se às áreas alpinas da Europa, e Wisconsin se refere à glaciação da América do Norte. Esta glaciação abrange o último período glacial inteiro, que começou cerca de 115.000 anos AP e terminou cerca de 11.700 anos AP. O UMG é uma fase dentro da glaciação Würm-Wisconsiana, correspondendo ao ponto de maior extensão das calotas de gelo, ocorrendo há aproximadamente 26.500 AP a 20.000 AP.

De acordo com a Comissão Estratigráfica Internacional (vide Anexo A), a fase que abrange o Último Máximo Glacial é a fase mais recente do Pleistoceno, denominada de Pleistoceno Superior. Este segmento corresponde à quarta e última subdivisão do Pleistoceno.

O Holoceno, por sua vez, teve início ao final da última grande glaciação (Würm), há cerca de 11.700 anos AP. Trata-se do período em que se reconhece o aparecimento do homem moderno e o desenvolvimento das primeiras grandes civilizações. As principais alterações climáticas ocorridas nesse intervalo caracterizam-se pelo início de uma fase interglacial em nosso planeta. Acredita-se que essa fase dure até hoje e deva terminar em algum momento nos próximos 1.000 anos (SUGUIO, 1999; SANT'ANNA NETO; NERY, 2005).

Assim, toda a história da humanidade até a presente data aconteceu numa época mais quente, também denominada interglacial, mesmo que pequenos avanços e recuos de fases mais frias tenham interferido nos climas terrestres (SANT'ANNA NETO; NERY, 2005). Concomitante a estas alterações climáticas, o nível relativo do mar também é fortemente influenciado por estes eventos, o que interfere diretamente nos níveis de base local, perturbando os processos sedimentares por todo o globo.

Apesar do Holoceno apresentar grandes eventos que perturbaram significativamente o clima global, o Pleistoceno Superior é marcado por períodos expressivos de avanços e recuos das geleiras, motivados por condições atmosféricas e oceanográficas que são objetos de estudos para as mudanças climáticas.

Para Loiola (2014), os humanos e todos seus parentes ancestrais do gênero Homo surgiram num período em que a Terra passava por grandes mudanças climáticas no Pleistoceno, ou seja, grandes oscilações no clima desde 2,5 Ma (milhões de anos). Segundo este autor, os humanos modernos se diferenciaram enquanto espécie na passagem do Pleistoceno Médio ao Superior, na África, sob a ocorrência das glaciações *Riss* e *Würm*, e dezenas de estágios glaciais e interglaciais abruptos globais observados nos eventos denominados

Dansgaard–Oeschger (D-O) e *Heinrich* (H). Sobre estes eventos, um breve resumo será discorrido.

2.1.4 Os eventos de mudanças climáticas abruptas do Quaternário

Os eventos de mudanças climáticas abruptas ocorridos no final do Pleistoceno e início do Holoceno ainda não são bem compreendidos, mas são importantes elementos das complexas interações entre o gelo, atmosfera e oceanos para os padrões climáticos globais (STRIKIS et al., 2011; CUSTÓDIO, 2022). Na sequência temporal dos eventos mais antigos para os mais recentes (Tabela 2.2), temos:

- Eventos Heinrich (H): Ocorreram no Atlântico Norte durante o período glacial, com grandes descargas de icebergs a cada 7.000-10.000 anos, afetando a salinidade e a circulação oceânica (AMOC - *Atlantic Meridional Overtuning Circulation*), alterando os padrões climáticos.

- Eventos de Dansgaard-Oeschger (D-O): Oscilações rápidas de aquecimento e resfriamento durante a última Era Glacial, a cada 1.500 anos. Causados por flutuações na circulação oceânica e mudanças na salinidade devido ao derretimento de gelo. Impactaram significativamente os climas globais e ecossistemas, especialmente no hemisfério Norte.

- Bølling-Alerød (BA): Período de aquecimento rápido entre aproximadamente 14.700 AP a 12.900 anos AP, com aumento das temperaturas e degelo de grandes mantos glaciais. Marcado por um aquecimento significativo nas regiões do Atlântico Norte e Europa, seguido pelo evento Younger Dryas.

- Younger Dryas (YD): Entre 12.900 AP a 11.700 anos AP, um resfriamento abrupto no Atlântico Norte, após um período de aquecimento. Possivelmente causado pela desaceleração da circulação termohalina devido ao derretimento de geleiras, ou por um impacto de asteroide (FIRESTONE et al., 2007). Não há um consenso sobre as causas do YD.

- Eventos de Bond: Ocorreram nos últimos 10.000 anos AP, a cada 1.000-1.500 anos, causando resfriamento e intensificação das tempestades no Atlântico Norte. Possíveis causas incluem variações na atividade solar e mudanças na

circulação oceânica, influenciando eventos históricos e as primeiras migrações humanas.

Tabela 2.2 – Eventos e períodos de alterações climáticas abruptas ocorridos nos últimos 21.000 anos AP.

Períodos	**Milhares de Anos (AP)**
Heinrich Stadial 1 (HI)	19.00 - 14.64
Bølling-Alerød (BA)	14.67 - 13.00
Younger Dryas (YD)	12.90 - 11.70
Holoceno Inicial (IH)	11.70 - 8.2
Holoceno Médio (MH)	8.2 - 4.2
Holoceno Tardio (LH)	4.2 - 0.0

Fonte: Adaptado de Custodio (2022).

2.2 O Cerrado e suas diversas origens

Como e quando se originou o cerrado? Eis um questionamento tão intrigante quanto o próprio objeto de estudo, pois afinal o que é o cerrado? Estas indagações povoaram as mentes de ilustres pesquisadores naturalistas desde que o Brasil passou a ser colonizado. No decorrer de décadas, o debate acadêmico se intensificou em várias correntes, mas paulatinamente, através do avanço de vários métodos de investigação científica, algumas teorias começam a convergir para um consenso.

Para a pergunta sobre “o que é o cerrado?” a literatura apresenta várias definições como resposta. Pode ser um bioma, tipos de vegetação ou formas de vegetação. E, antes de mais nada, o cerrado, enquanto bioma, é um mosaico vegetacional. Talvez pela complexidade da definição do que seja o cerrado, tenha advindo, igualmente, muitas polêmicas sobre sua origem.

Como afirmam Ribeiro e Walter (1998), o termo “cerrado”, de origem espanhola (fechado), tem sido usado tanto para designar tipos de vegetação (fisionomia, flora e ambiente) quanto para definir formas de vegetação (fisionomia, estrutura, formas de crescimento). Disso, se depreende na literatura, muitas vezes, a

utilização do termo “fitofisionomia”, abrangendo o tipo e a forma ao mesmo tempo.

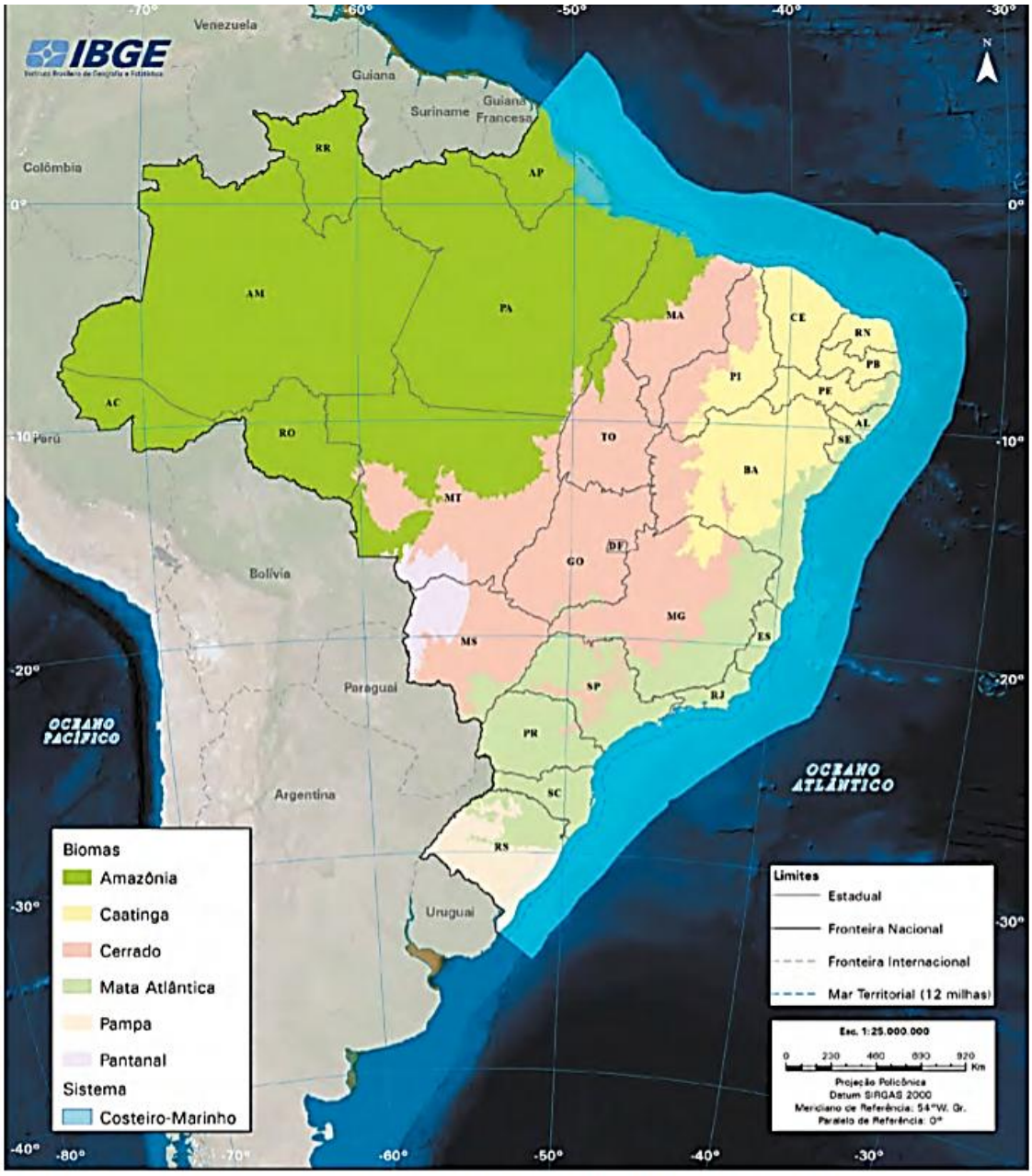
De acordo com Leite (2004), o Cerrado é o mais importante bioma campestre (estacional) brasileiro, tanto por sua extensão territorial quanto pela complexidade, ecologia, importância social e econômica. O bioma cerrado, do modo como está cartografado oficialmente pelo IBGE (Figura 2.5) com detalhamento em 1:250.000, estende-se aproximadamente entre a Amazônia (a oeste) e a Caatinga e a Mata Atlântica (a leste); entre os lençóis maranhenses (ao norte) e o pantanal mato-grossense (ao sul/sudoeste).

O cerrado incorpora um conjunto de tipologias vegetais de características significativamente similares, congêneres ou aproximadas em função das quais é passível de particularização como um território contínuo (LEITE, 2004; IBGE, 2012).

Para Ribeiro e Walter (1998), o emprego do termo "Cerrado" evoluiu para três acepções principais, as quais demandam distinção clara. A primeira refere-se ao bioma Cerrado (grafado com inicial maiúscula), predominante no Brasil Central, que não deve ser designado no plural ("Cerrados"), por tratar-se de uma unidade biogeográfica única. Quando mencionada a "região do Cerrado", alude-se à área geográfica deste bioma. A segunda acepção, cerrado sentido amplo (*lato sensu*), abrange as formações savânicas e campestres do bioma — incluindo o cerradão e o campo limpo (COUTINHO, 1978; EITEN, 1994) —, definido essencialmente pela composição florística e fisionomia (formas de crescimento), sem ênfase na estrutura. Alguns autores utilizam "cerrados" neste contexto específico, não confundível com o bioma. A terceira acepção, segundo Ribeiro e Walter (1998), temos o cerrado sentido restrito (*stricto sensu*) designado por uma fitofisionomia savânica específica, caracterizada por critérios florísticos, fisionômicos e estruturais, sendo emblemática do bioma Cerrado. Outros termos (como campo cerrado, cerrado denso ou cerrado interfluvial) constituem variantes ou subdivisões das acepções anteriores, principalmente do sentido amplo, refletindo tipologias de vegetação conforme autores ou escalas de análise.

Para efeito desta pesquisa, o conceito de cerrado está englobando a perspectiva de Pivello (2013), como uma fitofisionomia que apresenta uma vegetação sazonal com dois componentes: um estrato herbáceo-subarbustivo e outro arbustivo-arbóreo, sem a formação de um dossel contínuo.

Figura 2.5 - Mapa de Biomas e Sistema Costeiro-Marinho do Brasil.



Fonte: Adaptada de IBGE (2019).

2.2.1 O debate acadêmico sobre as origens do cerrado

Vários autores descrevem os embates científicos no entendimento sobre a origem e a evolução do cerrado. Lucena e Araújo (2020) descrevem um histórico interessante sobre o debate acadêmico das origens do cerrado, desde as primeiras incursões científicas europeias ao Brasil imperial, enfocando os trabalhos dos dinamarqueses Peter Wilhelm Lund (1801-1880) e de seu secretário, Eugenius Warming (1841-1924), nas primordiais descrições científicas sobre o cerrado.

Ribeiro e Walter (1998) atribuem a Warming quem primeiro indicou o clima como fator determinante das formações savânicas e campestre do cerrado, onde a deficiência hídrica no período seco seria o principal fator limitante (WARMING, 1908). De um modo geral, Walter e Ribeiro mencionam três vertentes para a origem das savanas da América Tropical: teorias climáticas, pelas quais a vegetação seria o resultado do clima, principalmente em função da limitação sazonal de água no período seco (estacionalidade); teorias bióticas, nas quais a vegetação seria o resultado de ação antrópica, principalmente pelo uso frequente do fogo; ou ainda resultante da atividade de outros agentes da biota como as formigas; e as teorias pedológicas, em que a vegetação seria dependente de aspectos edáficos e geológicos, como deficiências minerais (oligotrofismo), saturação por elementos como alumínio, diferenças de drenagem e profundidade dos solos.

Um aspecto importante a ressaltar é que diversos estudos respaldam uma vertente teórica de unificação de fatores bióticos e abióticos que podem ter levado a originar o cerrado. Muitos autores concordam que a interação entre esses fatores teria imposto uma forte pressão seletiva sobre as espécies de plantas do cerrado, sendo que o efeito desses fatores variou de acordo com as características ambientais, como: precipitação sazonal, baixa fertilidade do solo, incêndios e até mesmo as mudanças climáticas de períodos glaciais e interglaciais ocorridas no Quaternário. Também devem ser levados em consideração fatores espacialmente mais específicos, como características

hidrológicas e topográficas regionais (KAUFFMAN et al., 1994; OLIVEIRA FILHO; RATTER, 2002).

Pinheiro e Monteiro (2010) ressaltam que a redução na disponibilidade de umidade ambiental seria, de fato, o fator que desencadeou os mecanismos evolutivos responsáveis pelo estabelecimento dessa nova flora denominada savana. A secura neotropical, que atingiu seu pico durante a maior parte do período Quaternário, favoreceu a ocorrência de incêndios, que por sua vez contribuíram para o empobrecimento do solo e volatilização de nutrientes edáficos.

2.2.2 Aspectos das origens do cerrado e o fogo

Um dos aspectos mais característicos do cerrado é a frequente ocorrência de queimadas. Inúmeros trabalhos abordam os aspectos ecológicos do fogo no cerrado, principalmente nas espécies do estrato herbáceo-arbustivo (COUTINHO, 1977).

Simon et al. (2009) realizaram uma análise genética para fornecer novos *insights* sobre o papel do fogo na formação do bioma cerrado. Ao estudar a evolução de espécies do gênero Mimosa através da filogenia do DNA, os pesquisadores descobriram que, embora o cerrado tenha poucos gêneros endêmicos, existem espécies que se originaram exclusivamente dentro deste bioma. O objetivo da pesquisa era investigar a evolução das adaptações das plantas do cerrado ao fogo e obter uma perspectiva histórica das espécies nesse bioma.

O estudo de Simon et al. (2009) indicou que os incêndios naturais são um fator chave que contribui para que o cerrado seja a savana tropical com maior biodiversidade do mundo. Segundo Simon et al. (2009), a origem do cerrado coincidiu com um período de maior vulnerabilidade a incêndios naturais. Os pesquisadores propuseram que a origem do cerrado foi síncrona com a expansão de gramíneas C4 na região, impulsionada por mudanças climáticas ocorridas há aproximadamente 10 milhões de anos (Mioceno). Essas mudanças também facilitaram a expansão dos biomas de savana globalmente. Como essas gramíneas são suscetíveis ao fogo, elas tornaram as regiões de cerrado mais propensas a incêndios. Consequentemente, as plantas mais adaptadas a

ambientes secos e resistentes ao fogo tiveram maiores chances de sobrevivência.

O estudo de Simon et al. (2009) também sugeriu que a diversificação da maior parte da flora do cerrado começou há aproximadamente 4 milhões de anos (Plioceno). Biomas vizinhos, como a Amazônia, Caatinga, Pampas e Mata Atlântica, contribuíram para a formação do cerrado, compartilhando vários gêneros em comum. Essas descobertas implicam que o cerrado se formou *in situ* por meio de mudanças adaptativas recentes e frequentes para resistir ao fogo, e não pela dispersão de linhagens já adaptadas ao fogo. Apesar de utilizar apenas uma amostragem de DNA de 3 a 4% da flora do cerrado, Simon et al. (2009) trouxeram uma contribuição importante para a compreensão da história evolutiva das plantas do cerrado e sua adaptação ao fogo.

2.2.3 Teria o cerrado uma origem mais antiga?

A pesquisa de Aguiar et al. (2019) propõe uma hipótese de origem ainda mais antiga para o cerrado. Evidências sugerem que os primórdios do cerrado remontam ao Paleoceno. Baseando-se em um estudo filogenético de um clado de abelhas do gênero *Tapinotaspidini*, os autores identificaram grupos endêmicos de abelhas em áreas de savanas e outros tipos de vegetação aberta, além de áreas florestadas. A hipótese da pesquisa de Aguiar et al. (201) foi de que essas abelhas, originárias de biomas semelhantes à savana, mudaram o hospedeiro floral para aumentar a diversificação das espécies ao longo de sua história evolutiva. Utilizando ferramentas macroevolucionárias, os pesquisadores reconstruíram o habitat ancestral das abelhas e analisaram o efeito das mudanças no tipo de vegetação e na diversificação da espécie.

Os resultados indicam que as abelhas *Tapinotaspidini* são originárias de uma savana primitiva, tipo cerrado, no Paleoceno. A pesquisa também concluiu que as mudanças para áreas florestadas ocorreram pelo menos três vezes em um período de 30 milhões de anos (Ma), causando aumentos nas taxas de diversificação. A família Malpighiaceae foi identificada como hospedeiro floral ancestral, com mudanças para outras famílias ocorrendo no Eoceno e no Mioceno. Segundo os autores mencionados, essa ligação entre biomas,

hospedeiros florais e taxas de diversificação proporcionou *insights* adicionais sobre a evolução das abelhas e da flora associada na América do Sul.

Além dos resultados de Aguiar e equipe, outros estudos destacaram evidências de uma origem mais antiga para vegetações campestres. A mudança anatômica na elaboração da fotossíntese (anatomia de Kranz) permitiu que as plantas se diferenciassem em C3 ou C4, crucial para a adaptação a climas mais tropicais. Um estudo de Urban et al. (2010) demonstrou que gramíneas C4 já eram abundantes na paisagem europeia no início do Oligoceno, desafiando a ideia de que se tornaram comuns apenas no final do Mioceno. Essas gramíneas predominavam no sudoeste da Europa, especialmente no sul da Espanha, até que o resfriamento climático e a mudança no regime de chuvas, concentradas nos meses de inverno, reduziram sua abundância.

Jacobs et al. (1999) estudaram a origem dos ecossistemas dominados por gramíneas na América do Sul, mencionando fósseis de dentes de ungulados endêmicos (notoungulados) do final do Paleoceno. Esses animais tinham dentes de coroa alta, adaptados ao consumo de vegetação abrasiva, sugerindo uma adaptação a dietas de gramíneas desde o final do Paleoceno.

2.3 Processos paleoclimáticos e a dinâmica biogeográfica do Brasil, no Último Máximo Glacial

A Biogeografia é a ciência que estuda as mudanças ambientais que influenciam a distribuição de espécies nos biomas terrestres ao longo do tempo e tem contribuído significativamente, junto com a Paleoclimatologia, para a compreensão de como as espécies de formações vegetais interagem com os fatores ambientais na formação das diversas paisagens.

Arruda e Schaefer (2020) trazem uma abrangente abordagem sobre o “estado-da-arte” da dinâmica paleoclimática e biogeográfica brasileira no UMG e as controvérsias científicas sobre a dinâmica climática pretérita.

2.3.1 Último Máximo Glacial mais frio e seco ou mais frio e úmido?

As vegetações brasileiras são moldadas pela interação de fatores ambientais como geologia, clima e solo. As mudanças climáticas durante o Último Máximo

Glacial (UMG), que ocorreu há aproximadamente 18 mil anos, desempenharam um papel importante na configuração das vegetações no Brasil (ARRUDA; SCHAEFER, 2020).

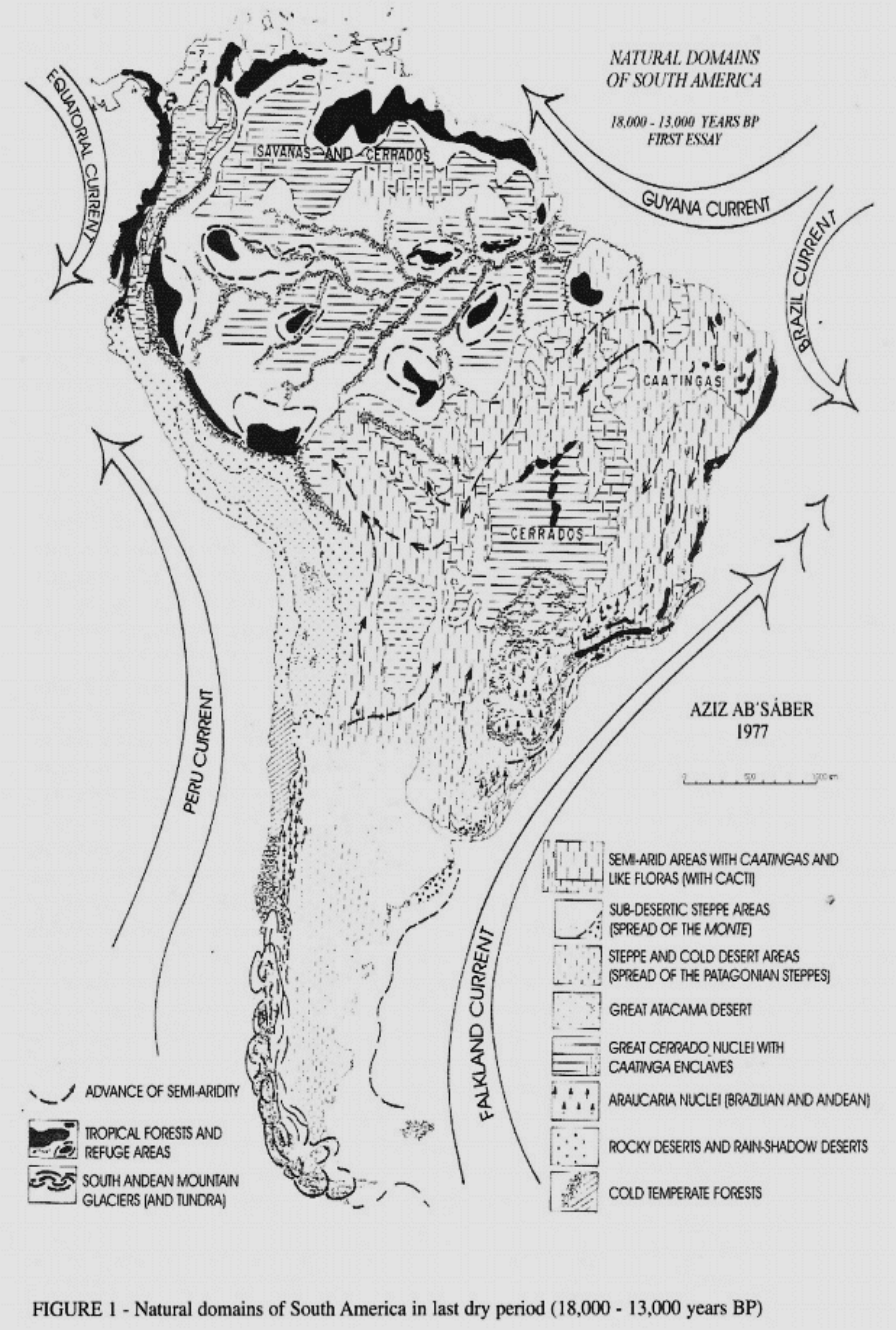
2.3.1.1 Hipótese do UMG mais frio e seco

Estudos geomorfológicos e palinológicos indicam que momentos glaciais podem ter sido frios e secos, com predominância de processos físicos de desintegração mecânica da rocha matriz e erosão superficial (BIGARELLA et al., 1994; AB'SABER, 2003). Este clima seco teria favorecido paisagens mais abertas, como caatingas e cerrados. A hipótese dos "Refúgios Amazônicos" de Haffer (1969) sugere que a floresta amazônica foi fragmentada durante o UMG, sendo parcialmente substituída por vegetações abertas.

Ab'Saber (1977) faz a primeira aproximação da cobertura vegetal brasileira no final do UMG, destacando a expansão do cerrado e da caatinga. Ao analisar a possível redução das precipitações ocorrida durante o UMG, Ab'Saber apresentou uma primeira estimativa da provável distribuição da vegetação brasileira ao final da última glaciação em resposta a um clima mais frio e seco.

Para Ab'Saber, há a expansão do cerrado sobre a floresta na região da Amazônia, conforme ilustrado na Figura 2.6, onde somente refúgios de floresta permaneceriam nas nascentes dos principais rios (conforme também proposto por Haffer, em 1969). Além disso, Ab'Saber aponta a expansão da caatinga sobre o cerrado e uma parte da Mata Atlântica no Sudeste, o avanço mais ao norte das matas de araucárias e a expansão da vegetação campestre do Sul (pampas gaúchos). Em síntese, ele propõe que, no UMG, o clima favoreceu uma expansão de fitofisionomias sazonais mais abertas em detrimento das formações florestais. (ARRUDA; SCHAEFER, 2020).

Figura 2.6 - Domínios naturais da América do Sul.



Fonte: Adaptado de Ab'Saber (1977).

A configuração dos domínios vegetacionais brasileiros, segundo Ab'Saber, está ligada ao fortalecimento das correntes marinhas frias originárias da Antártica. Essas correntes gerariam ventos frios e secos para o interior do continente, além de diminuir os ventos alísios, resultando na redução da precipitação próxima à região equatorial.

2.3.1.2 Hipótese do UMG mais frio e úmido

Segundo Arruda e Schaefer (2020), pesquisas indicam que durante o Último Máximo Glacial (UMG), o clima foi frio, mas não tão seco, permitindo uma maior cobertura florestal. Esta hipótese é suportada por achados de pólen de espécies andinas na Amazônia, sugerindo que a bacia amazônica sempre foi florestada (WANG et al., 2017; COLINVAUX et al., 1996; 2000). Arruda et al. (2018) propõem uma hipótese intermediária, onde o UMG foi frio e úmido, favorável ao estabelecimento de uma floresta similar à atlântica costeira, o que explicaria, desta forma, a presença de pólen de florestas andinas.

De acordo com essa hipótese, a umidade durante o UMG foi suficiente para permitir a migração e distribuição de espécies andinas para o interior da Amazônia, o que gerou polêmica em relação ao modelo de vegetação pré-estabelecido para o UMG. A discussão sobre o clima no UMG ainda persiste, especialmente no que diz respeito à precipitação. Há um consenso de que houve um esfriamento uniforme de aproximadamente 5°C nas Américas, alcançando até 5,4°C no Brasil Tropical (STUTE et al., 1995).

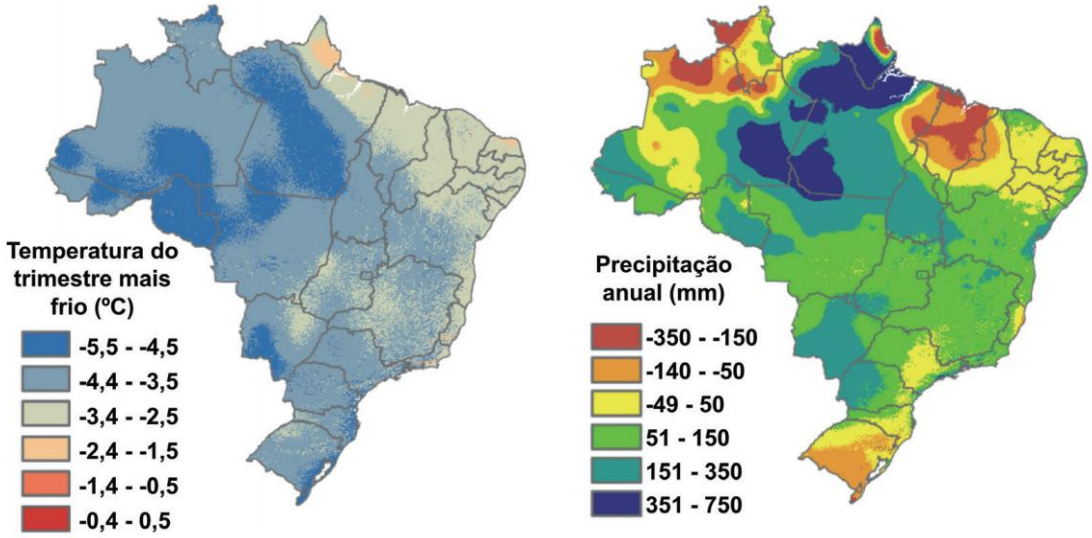
Embora haja entendimento convergente quanto ao esfriamento, os fenômenos complexos que regem o sistema de precipitação na América do Sul, especialmente na bacia amazônica, são menos claros. Diversos estudos sugerem a ocorrência de um “efeito gangorra” ou fenômeno de antifase entre o sistema de precipitação na porção ocidental da Amazônia e na porção oriental amazônica e nordeste brasileiro, durante o verão austral no UMG, juntamente com o deslocamento da Zona de Convergência Intertropical (ZCIT) para latitudes mais altas (CRUZ et al., 2009).

Atualmente, sabe-se que a precipitação sobre a América do Sul tropical é amplamente controlada pelos movimentos sazonais da ZCIT. Durante o verão, a ZCIT se desloca para o sul devido ao aquecimento do continente, aumentando a intensidade das chuvas no interior do Brasil.

Os modelos climáticos globais utilizam ciclos hidrológicos, modificações na concentração de CO_2 atmosférico, temperatura da superfície terrestre e do mar, interações de *feedback* e forçantes externas para reconstruir condições

climáticas passadas. Os modelos atuais sugerem que o clima no UMG foi caracterizado por condições mais frias e úmidas em relação aos dias atuais, indicando que o clima amazônico durante o UMG foi complexo e mais úmido do que atualmente (Figura 2.7). Esse aumento na umidade pode ser atribuído às forçantes orbitais e ao deslocamento da ZCIT.

Figura 2.7 - Diferença climática entre o UMG (20 mil anos AP) e a atualidade, considerando a temperatura e a precipitação.



Fonte: Adaptado de Arruda e Schaefer (2020).

Em resumo, Arruda e Schaefer (*op. cit.*) concluem que foi o aumento das temperaturas, e não sua diminuição, aliado à redução das precipitações no período pós-glacial, que favoreceu a expansão das vegetações sazonais, onde as condições de solo eram propícias. Além disso, destacam que durante o UMG houve uma redução das chuvas no sul do Brasil. Esta condição favoreceu o avanço dos campos sulinos e o direcionamento das florestas de araucária mais para o norte (BEHLING; HOOGHIEMSTRA, 2001; BEHLING, 2002; BEHLING et al., 2007), chegando até a região do Triângulo Mineiro (LEDRU et al., 1996). A recomposição das áreas de florestas e o recuo dos campos sulinos ocorreram após o UMG.

2.3.2 O registro de pólen no paleoambiente e a evolução do cerrado

Ledru (2002) estudou registros palinológicos do cerrado, indicando que a composição florística da vegetação variou ao longo do tempo e espaço, conforme as condições ambientais locais. O registro mais antigo é datado de cerca de 32.000 anos. A fitofisionomia atual do cerrado se estabeleceu há cerca de 7.000 anos no Brasil central e 10.000 anos no norte do Brasil, como consequência do aumento da sazonalidade e da temperatura. A influência do homem e o aumento das frequências de fogo contribuíram para essas mudanças. Em períodos mais secos e/ou frios, áreas de floresta úmida se conectaram através de florestas de galeria, contribuindo para diferentes fitofisionomias do cerrado observadas atualmente.

Barberi (2001) estudou sedimentos na Lagoa Bonita (DF) e concluiu que a evolução do cerrado e a vegetação neotropical estão relacionadas às histórias geológica e climática pré-pleistocênicas da América do Sul, com influência da antiga província florística de Gondwana. Ele identificou dois intervalos distintos na evolução paleoambiental da região, com um período seco intermediário. O Pleniglacial superior (28 a 13 mil anos AP) foi caracterizado por um clima mais úmido e frio, com predominância de elementos arbóreos e palustres, incluindo a palmeira buriti (*Mauritia flexuosa*). Já o Glacial tardio (13 a 10 mil anos AP) teve predomínio de cerrado, indicando um clima com duas estações. A vegetação da vereda, identificada a partir da ocorrência de *Mauritia* na parte superior dos sedimentos, remonta a aproximadamente 6.300 ± 40 AP.

Strikis et al. (2011) estudaram as variações de pluviosidade durante o glacial tardio e o Holoceno, observando que a duração dos eventos abruptos de pluviosidade foi mais longa no Holoceno Inicial e Médio, mas mais curta no Holoceno Superior. As mudanças climáticas ao longo do Holoceno foram associadas à redução da cobertura de gelo no Hemisfério Norte, que levou à diminuição da circulação termohalina, caracterizando os eventos Bond.

Custódio (2020) estudou a variabilidade do Sistema de Monções da América do Sul (SMAS) nos últimos 21.000 anos, utilizando simulações do TraCE-21k (*Transient Climate Evolution over the last 21.000 years*). O autor encontrou que

a precipitação associada ao SMAS e ao posicionamento da Zona de Convergência do Atlântico (A-ZCIT) foi influenciada pelos pulsos de água doce do degelo do Hemisfério Norte, especialmente entre o *Heinrich Stadial* 1 (H1) e o *Younger Dryas* (YD). Durante os períodos frios do Hemisfério Norte, houve mudanças no regime de precipitação devido ao deslocamento da ZCIT para o sul, afetando o nordeste brasileiro e a porção central da Bacia Amazônica.

Cassino et al. (2017) estudaram o registro de pólen fóssil de um núcleo de turfa do Quaternário tardio, coletado no pântano de palmeiras São José, na região noroeste de Minas Gerais (Brasil Central). O estudo comparou análogos modernos com amostras fósseis de pólen. Os resultados sugerem que entre 15.700 e 15.400 anos calibrados AP, a palmeira buriti (*Mauritia flexuosa*) estava ausente, indicando um clima mais frio.

Por volta de 15.200 anos calibrados AP, o aumento da ocorrência de buriti e o desenvolvimento do cerrado stricto sensu indicaram um clima mais quente. No final do Pleistoceno Superior e durante a transição Pleistoceno-Holoceno, a zona central do pântano de palmeiras se expandiu, e um denso cerrado stricto sensu se desenvolveu, sugerindo um clima quente e úmido.

No final do Holoceno Inferior, um clima mais seco é indicado pelo encolhimento da palmeira de pântano e uma vegetação regional mais aberta. Até cerca de 4.400 anos calibrados AP, um clima mais úmido é indicado pela expansão máxima da vereda São José em cerca de 6.500 anos AP. Após 3.500 anos calibrados AP, o predomínio do Campo Sujo (vegetação herbáceo-arbustiva) indica o retorno de condições relativamente secas. Nos últimos dois milênios, a diminuição da zona de umidade do pântano de palmeiras sugere uma intensificação das condições mais secas.

A tendência de menor precipitação no final do Holoceno, enquanto o Holoceno Médio foi predominantemente mais úmido do que o presente, é diferente do padrão observado no sul e oeste do cerrado, mas está de acordo com os dados de paleoclima do Nordeste do Brasil. Na região do sítio de São José, temperaturas semelhantes às atuais parecem ter prevalecido desde o

Pleistoceno Superior, mas grandes variações na disponibilidade de umidade foram registradas durante o Holoceno.

Importante ainda ressaltar que os autores comentam sobre a oscilação na quantidade de precipitação durante o Holoceno. No entendimento dos autores, esta oscilação é corroborada por um registro isotópico de oxigênio de um espeleotema da caverna da Lapa Grande (STRÍKIS et al., 2011), localizada a cerca de 300 km ao norte da vereda de São José, que mostrou a ocorrência de doze eventos de aumento da precipitação ao longo do Holoceno, intercalado com fases menos úmidas. A cronologia desses eventos não é diretamente comparável às fases de expansão e retração do pântano de palmeiras de São José e de outro sítio estudado, denominado de Laçador, devido à baixa resolução temporal dos dados palinológicos, mas este registro paleoclimático confirma variações periódicas na quantidade de precipitação nesta região durante o Holoceno.

Cassino et al. (2017) também realizam uma análise comparativa com todos os sítios que contém pesquisas de registros palinológicos. A maioria deles, em área de cerrado. Em alguns sítios foram encontrados sincronicidades de eventos com a área de estudo do sítio São José, e em outros não. Isso está de acordo com os trabalhos de Barberi e Lima (2008), sobre o não sincronismo dos efeitos paleoclimáticos na dinâmica da vegetação.

Os registros pontuais dos testemunhos palinológicos pretéritos necessitam de um volume maior de elementos espacialmente dispersos para uma análise evolutiva mais clara do ambiente e dos sistemas climáticos pretéritos.

2.4 Considerações da revisão para os objetivos da pesquisa.

Após a revisão teórica sobre a evolução do cerrado a partir da última glaciação, conclui-se que ainda não há um consenso inequívoco sobre a origem do cerrado, e que os estudos precisam continuar a fim de elucidar antigas e novas lacunas.

Com relação aos climas passados e suas influências na evolução dos ecossistemas, no Último Máximo Glacial (UMG) houve um resfriamento global com impactos na cobertura vegetal. A Teoria de Milankovitch explica a influência

das variações orbitais na radiação solar, afetando a distribuição de ecossistemas. Entretanto, a circulação oceânica e eventos climáticos abruptos, como os eventos de Heinrich, Dansgaard-Oeschger, Younger Dryas e Bond, também tiveram papel fundamental na evolução da biota terrestre.

A origem do cerrado ainda é amplamente debatida. Diferentes hipóteses sugerem que sua formação se deu por fatores climáticos, pedológicos e antrópicos. Estudos genéticos indicam que o cerrado pode ter se originado há cerca de 10 milhões de anos, com influência de gramíneas C4 e do fogo. Outros estudos sugerem uma origem ainda mais remota, no Paleoceno, baseada em registros de DNA de especializações de abelhas e da dentição de herbívoros extintos.

Durante o UMG, a dinâmica climática influenciou diretamente a distribuição das fitofisionomias vegetais no Brasil. A hipótese dos refúgios florestais sugere que a floresta Amazônica se fragmentou em pequenos núcleos devido à aridez. Contudo, evidências recentes sugerem um UMG mais úmido na região, desafiando essa hipótese. Estudos de pólen confirmam a expansão de florestas de araucárias para o norte e a substituição gradual de vegetação sazonal por florestas tropicais após o UMG.

As análises de pólen mostram que a vegetação do cerrado se consolidou entre 7.000 e 10.000 anos AP. Testemunhos de sedimentos indicam oscilações entre climas secos e úmidos, correlacionando-se às mudanças na intensidade do Sistema de Monções da América do Sul (CASSINO et al., 2020). Modelagens climáticas sugerem um Holoceno caracterizado por flutuações de umidade, influenciadas pela Temperatura da Superfície do Mar e a migração da Zona de Convergência Intertropical.

Os estudos paleoclimáticos e paleobiogeográficos fornecem alguns direcionamentos sobre os fatores ambientais limitantes e/ou favoráveis à ocorrência do cerrado através do tempo.

Barberi e Lima (2008) ressaltam que os fenômenos da dinâmica climática pretérita não são sincrônicos, sendo possível constatar diferenças na resposta de cada região a estas alterações, havendo influência das características locais.

Para os autores, a interação entre os diferentes parâmetros físicos, como latitude, altitude, condicionamento geológico, compartimentação geomorfológica local e regional, tipo de solo, clima regional e as variações de clima local, além dos parâmetros biológicos, representados principalmente pela cobertura vegetal, ocorrência e intensidade de queimadas e mais recentemente da ação humana, são fundamentais para estabelecer a resposta de cada área às modificações paleoclimáticas. Essa influência dos parâmetros locais parece se acentuar principalmente quando as condições climáticas não são tão extremadas quanto aquelas que ocorrem durante os períodos glaciais (BARBERI; LIMA, 2008).

Muitos fatores ambientais modulam a ocorrência do cerrado e é esperado que a diversidade de fitofisonomias do cerrado responda aos parâmetros físicos e bióticos locais, associado às modificações climáticas e geológicas, tanto de caráter global como local, quando estudadas sob uma abordagem de modelos e padrões computacionais.

3 FUNDAMENTAÇÃO TEÓRICA DO DELINEAMENTO EXPERIMENTAL

Nesta seção, será discorrida a fundamentação teórica que embasou a execução do delineamento experimental. Serão descritos conceitos tanto de técnicas como de *softwares* necessários para os objetivos da pesquisa[2].

A explosão de dados disponíveis em diversas áreas do conhecimento impulsionou o desenvolvimento de técnicas avançadas para sua análise e interpretação. Entre essas técnicas, destacam-se *Machine Learning* (Aprendizado de Máquina) e Mineração de Dados, que permitem a extração de padrões, a previsão de eventos e a modelagem de fenômenos complexos.

3.1 Fundamentação teórica de machine learning, mineração de dados e suas aplicações

O termo “machine learning” (“aprendizado de máquina”, em português) foi criado por Arthur Samuel, em 1959, como "o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados” (SAMUEL, 1959). Em sua publicação seminal, o autor impulsionou os primórdios da inteligência artificial através do aprendizado computacional para uma partida de jogo de damas.

O aprendizado de máquina pode ser definido como um campo da inteligência artificial que se baseia no desenvolvimento de algoritmos capazes de aprender a partir de dados e tomar decisões sem necessidade de programação explícita. Os algoritmos de aprendizado de máquina podem ser classificados em três categorias principais:

a) Aprendizado supervisionado: o modelo é treinado com dados rotulados, ou seja, conhece-se a relação entre variáveis de entrada e saída. Exemplos incluem classificadores baseados em árvores de decisão, redes bayesianas e redes neurais;

[2] No final do documento, há um glossário com alguns termos que são utilizados frequentemente durante o trabalho e que dizem respeito ao universo do *Machine Learning*.

b) Aprendizado não supervisionado: não há rótulos nos dados, e o modelo precisa identificar padrões por si só. Técnicas como análise de agrupamento (*clustering*) e regras de associação pertencem a essa categoria;

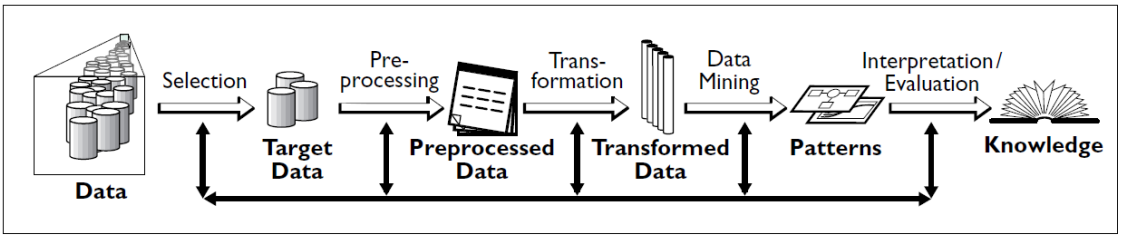
c) Aprendizado por reforço: o modelo aprende interagindo com um ambiente e recebendo recompensas ou punições por suas ações, buscando maximizar um retorno cumulativo ao longo do tempo.

A Mineração de Dados (*Data Mining*, em inglês) é um processo de descoberta de conhecimento em bases de dados extensas e complexas. Utiliza algoritmos de aprendizado de máquina para encontrar padrões ocultos, relações e tendências nos dados. São amplamente utilizadas nesta área, técnicas como classificação, regressão, segmentação e descoberta de regras de associação (HAN et al., 2012).

3.2 O método KDD: descoberta de conhecimento em bases de dados

Knowledge Discovery in Databases (KDD) é um processo estruturado que visa transformar grandes volumes de dados brutos em conhecimento útil (FAYYAD et al., 1996). A Figura 3.1 mostra um esquema do processo KDD.

Figura 3.1 - Esquema do processo KDD (Knowledge Discovery in Dataset).



Fonte: Adaptado de Fayyad et al. (1996).

O método compreende cinco etapas principais:

a) Seleção dos dados: identificação das fontes e variáveis relevantes para análise;

b) Pré-processamento: limpeza, tratamento de valores ausentes e normalização para garantir a qualidade dos dados;

c) Transformação: engenharia de atributos e redução de dimensionalidade para otimização dos modelos;
d) Mineração de Dados: aplicação de algoritmos para extrair padrões e relações ocultas;
e) Interpretação e Avaliação: validação dos modelos gerados e extração de conhecimento relevante.

O método KDD é considerado um arcabouço metodológico que engloba a Mineração de Dados e o *Machine Learning*, garantindo que a extração de informações relevantes seja feita de forma sistemática e interpretável.

3.3 QGIS

O QGIS (*Quantum Geographic Information System*) é um software de código aberto e gratuito para Sistemas de Informação Geográfica (SIG), desenvolvido por uma comunidade mundialmente interativa. É um software utilizado para visualizar, editar, analisar e publicar dados geoespaciais. Compatível com Windows, macOS e Linux, ele oferece suporte a diversos formatos de dados, como Shapefile, GeoJSON, GeoTIFF e bancos de dados espaciais (PostGIS, SpatiaLite). Sua interface intuitiva e modular permite a integração com ferramentas avançadas além de contar com uma vasta biblioteca de plugins para expandir suas funcionalidades. Ideal para aplicações em geoprocessamento, cartografia e análises ambientais, o QGIS é amplamente adotado por profissionais, pesquisadores e instituições devido à sua flexibilidade e não apresentar custos (QGIS.ORG, 2023).

Desenvolvido por uma comunidade global ativa, o QGIS é constantemente atualizado, garantindo recursos modernos como criação de mapas temáticos, processamento em lote e integração com serviços web (WMS, WFS). Sua licença open-source (GPL) permite livre acesso e personalização, tornando-o uma alternativa robusta para projetos de planejamento urbano, agricultura de precisão, conservação ambiental e outras atividades geoespaciais (QGIS.ORG, 2023).

3.4 WEKA

O WEKA (*Waikato Environment for Knowledge Analysis*) é uma plataforma de código aberto para aprendizado de máquina e mineração de dados, desenvolvida pela Universidade de Waikato, na Nova Zelândia. Escrito na linguagem Java, o *software* oferece uma coleção abrangente de algoritmos prontos para uso, incluindo classificadores (como J48 e Random Forest), métodos de *clustering*, regras de associação e técnicas de pré-processamento de dados (WITTEN et al., 2016). Sua interface gráfica intuitiva (Figura 3.2) permite que usuários sem programação avançada apliquem esses algoritmos a conjuntos de dados. Apresenta também uma API Java, que possibilita integração em projetos mais complexos.

O WEKA é amplamente utilizado em pesquisas acadêmicas e aplicações práticas, como análise preditiva, bioinformática e processamento de textos, graças à sua versatilidade e suporte a formatos como ARFF, csv (*Comma-Separated Values,* ou seja, valores separados por vírgula) e bancos de dados. Além disso, a plataforma inclui recursos como visualização de dados, seleção de atributos e validação cruzada, facilitando experimentos e comparações de modelos (WITTEN et al., 2016).

Figura 3.2 – Interface do WEKA, onde aparece a ave da Nova Zelândia que dá origem ao nome do aplicativo.



Fonte: Adaptado de Witten et al. (2016), Apêndice online.

3.5 Regras de associação e suas métricas

As **regras de associação** são técnicas de mineração de dados utilizadas para identificar relações frequentes entre itens dentro de um conjunto de dados. Elas são amplamente empregadas em análise de cestas de compras (*Market Basket Analysis*), bioinformática, e diversas outras áreas (HAN et al., 2012).

Um dos algoritmos mais conhecidos para extração de regras de associação é o Apriori, que opera descobrindo padrões frequentes e estabelecendo regras do tipo:

“Se (A e B) ocorrem juntos, então C também ocorre com uma determinada probabilidade”.

Para medir a qualidade dessas regras, são utilizadas diversas métricas, incluindo:

a) Suporte: Frequência com que um conjunto de itens aparece na base de dados.
b) Confiança: Probabilidade de encontrar o item B em transações que contenham o item A.
c) *Lift*: Mede a força da associação entre itens comparada ao acaso (valores acima de 1 indicam associação positiva).
d) *Conviction*: Mede a força da implicação entre dois itens. Quanto maior o valor, maior a certeza de que a regra é válida. Um valor de 1 indica independência total entre os itens.
e) *Leverage*: Avalia o impacto da associação entre dois itens em comparação ao que seria esperado por acaso. É calculado como a diferença entre o suporte observado e o suporte esperado. Valores positivos indicam associação positiva.

Essas métricas ajudam a selecionar as regras mais informativas e confiáveis dentro de um grande conjunto de dados.

3.6 Árvores de Decisão

Uma Árvore de Decisão é uma estrutura hierárquica onde cada nó interno representa uma decisão baseada em um atributo dos dados, e cada folha representa um resultado ou classe (QUINLAN, 1993).

- Interpretação: Árvores de decisão são fáceis de interpretar e visualizar. Cada caminho da raiz até uma folha representa uma regra de decisão clara.

- Desempenho: Árvores de decisão podem sofrer de *overfitting*, especialmente se forem muito profundas ou complexas. Isso significa que elas podem se ajustar muito bem aos dados de treinamento, mas não generalizam bem para novos dados.

- Velocidade: Treinar uma árvore de decisão é relativamente rápido, mas a eficiência depende do algoritmo de construção utilizado.

3.7 Random Forest

O algoritmo Random Forest (RF) é um conjunto de múltiplas árvores de decisão, combinadas para melhorar a precisão e a robustez do modelo. Cada árvore é construída a partir de um subconjunto aleatório dos dados de treinamento e dos atributos (BREIMAN, 2001).

- Interpretação: Random Forests são menos interpretáveis que árvores de decisão individuais, pois consistem em muitas árvores diferentes. No entanto, elas são mais robustas e menos suscetíveis a *overfitting*.

- Desempenho: Random Forests tendem a ter um desempenho melhor em termos de precisão e generalização em comparação com árvores de decisão individuais. A combinação de várias árvores reduz a variabilidade e melhora a robustez do modelo.

Treinar um classificador Random Forest é mais demorado do que treinar uma única árvore de decisão devido ao número de árvores que precisam ser construídas. No entanto, o treinamento pode ser paralelizado, ou seja, distribuído a carga de trabalho entre diferentes nós ou núcleos de processamento.

3.7.1 Diferenças principais entre Árvore de Decisão e Random Forest

- Complexidade: Árvores de decisão são simples e fáceis de entender, enquanto Random Forests são mais complexas e robustas.

- Interpretação: Árvores de decisão são interpretáveis, mas propensas a *overfitting*. Random Forests, embora menos interpretáveis, fornecem melhores previsões devido à agregação de múltiplas árvores.

- Desempenho: Random Forests geralmente superam árvores de decisão individuais em termos de precisão e capacidade de generalização.

3.8 Redes bayesianas

São estruturas gráficas que modelam relações de dependência probabilística entre variáveis. São baseadas na Teoria de Bayes e utilizam probabilidades condicionais para realizar inferências sobre o sistema (PEARL, 1988).

Os componentes de uma rede bayesiana, são:

- **Nós**: Representam as variáveis aleatórias.

- **Arestas**: Indicam influência ou dependência direta entre os nós.

- **Tabelas de probabilidade**: Cada nó contém uma tabela com probabilidades condicionais baseadas em seus pais (variáveis predecessoras).

3.8.1 Vantagens de redes bayesianas

Redes Bayesianas podem aprender relacionamentos causais e serem usadas para entendimento sobre o domínio de um problema e predizer consequências de uma intervenção.

Uma rede bayesiana é uma representação ideal para combinar conhecimento a priori e conhecimento proveniente dos dados, pois fornece uma abordagem apropriada para evitar *data overfitting*. Também são versáteis para raciocinar sobre um problema, pois são eficientes para construir modelos de domínios com incerteza inerente.

3.9 Métricas de desempenho dos algoritmos classificadores

Ao executar testes de classificação em algoritmos de aprendizado de máquina, algumas métricas precisam ser avaliadas sobre o desempenho da classificação pelo algoritmo utilizado. Entre as métricas, as mais utilizadas são:

a) Acurácia: Proporção de previsões corretas em relação ao total de previsões. É uma métrica intuitiva, mas pode ser enganosa em conjuntos de dados desbalanceados.
b) Precisão (*Precision*): Proporção de previsões positivas corretas (verdadeiros positivos) em relação ao total de previsões positivas (verdadeiros positivos + falsos positivos). É importante quando o custo dos falsos positivos é alto.
c) Recall (Sensibilidade): Proporção de verdadeiros positivos em relação ao total de casos positivos reais (verdadeiros positivos + falsos negativos). Útil quando o custo dos falsos negativos é significativo.
d) F1-Score: Combina precisão e recall em uma única métrica através de sua média harmônica. É útil quando queremos encontrar um equilíbrio entre precisão e recall.
e) Curva ROC e AUC: A curva Característica de Operação do Receptor (ROC) mostra a performance do modelo em todos os limiares de classificação. A área sob a curva ROC (AUC) fornece uma medida agregada de performance em todos os limiares de classificação.
f) Matriz de Confusão: Apresenta uma tabela que compara as previsões do modelo com os rótulos verdadeiros, dividida em verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN).

No Capítulo 5 serão detalhados os Materiais e Métodos da pesquisa e como foram utilizados os algoritmos e suas métricas teoricamente descritos neste capítulo.

4 INFORMAÇÕES AMBIENTAIS EM REPOSITÓRIOS INSTITUCIONAIS SOBRE O CERRADO UTILIZADAS NA PESQUISA

Existem uma diversidade de informações ambientais que estão depositadas em repositórios institucionais, no Brasil. A expressão "repositórios institucionais", utilizada neste documento está sendo usada para se referir às bases de dados, sistemas ou plataformas que armazenam informações disponíveis em instituições como, por exemplo, a INDE (Infraestrutura Nacional de Dados Espaciais), o SGB (Serviço Geológico do Brasil), o INPE (Instituto Nacional de Pesquisas Espaciais), o IBGE (Instituto Brasileiro de Geografia e Estatística), entre outros. Em sua maioria, os repositórios citados contêm informações geospaciais.

O termo descreve bem o papel desses sistemas, que geralmente centralizam, organizam e disponibilizam informações criadas e mantidas pelas instituições, frequentemente com foco em transparência, pesquisa e acesso público. Eles funcionam como verdadeiros "repositórios" de dados e conhecimento.

Nesta pesquisa foram usados os dados dos repositórios das instituições federais: IBGE, EMBRAPA e CPRM. A Tabela 4.1 mostra um breve resumo dos repositórios e quais informações foram acessadas para os objetivos do estudo.

A Figura 4.1 mostra a interface de acesso a uma das Base de Dados, a do BDiA, de onde foram acessados os dados, em escala 1:250.000, para a pesquisa.

Figura 4.1 – Portal do BDiA do IBGE.



Fonte: IBGE (2023).

Tabela 4.1 – Dados ambientais, instituição do repositório, escala/resolução espacial e endereço da url.

DADOS AMBIENTAIS	repositório/ Instituição	Escala ou resolução espacial	URL
GEOLOGIA	BDIA/IBGE	1:250.000	https://bdiaweb.ibge.gov.br
GEOMORFOLOGIA	BDIA/IBGE	1:250.000	https://bdiaweb.ibge.gov.br
SOLOS	BDIA/IBGE	1:250.000	https://bdiaweb.ibge.gov.br
VEGETAÇÃO	BDIA/IBGE	1:250.000	https://bdiaweb.ibge.gov.br
GRADE ESTATÍSTICA	IBGE	1 km x 1 km	https://www.ibge.gov.br/geociencias/downloads-geociencias.html
REDE HIDROGRÁFICA	IBGE	1:250.000	https://www.ibge.gov.br/geociencias/downloads-geociencias.html
DISPONIBILIDADE DE ÁGUA NO SUBSOLO	ÁGUA NO SOLO/ EMBRAPA	1:250.000	https://geoinfo.dados.embrapa.br
DECLIVIDADE	CPRM/SGB	1:250.000	https://www.sgb.gov.br/mapa-de-declividade -em-percentual-do-relevo-brasileiro
DADOS CLIMATOLÓGICOS	IBGE	1:5.000.000	https://www.ibge.gov.br/geociencias/downloads-geociencias.html

Fonte: Elaborado pela autora.

4.1 Informações ambientais do BDiA

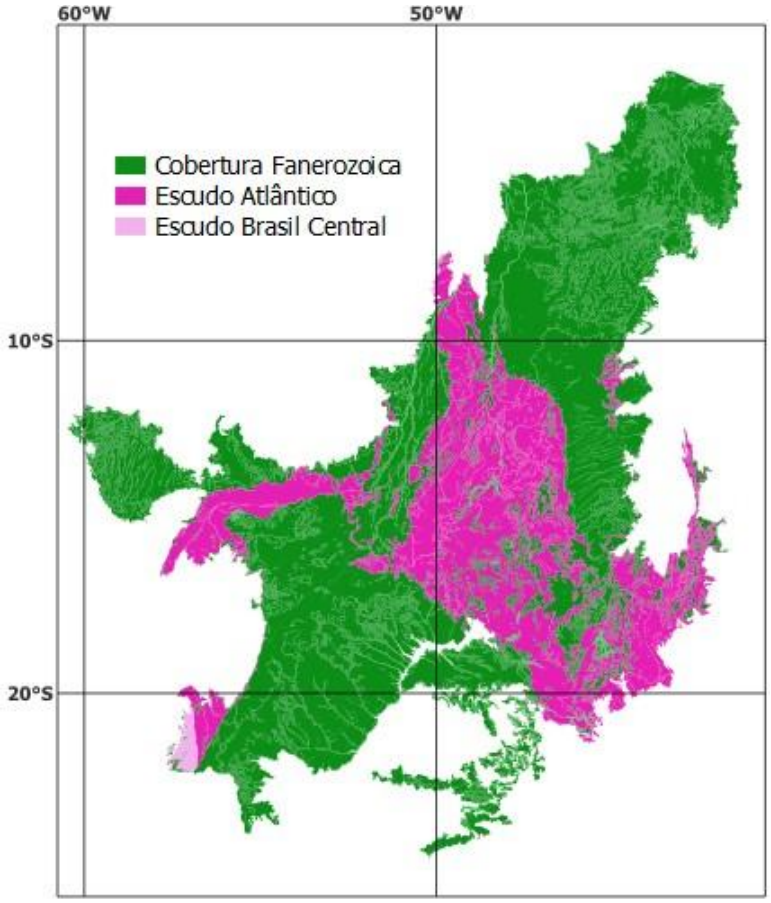
O BDiA é um sistema de informações ambientais para o Brasil, disponível à consulta pública através de uma aplicação web (disponível em https://bdiaweb.ibge.gov.br) que reúne a coleção de bases temáticas contínuas dos mapeamentos de recursos naturais do País dos temas Geologia, Geomorfologia, Pedologia e Vegetação, em escala 1:250.000, produzidos pelo IBGE. A plataforma tem como objetivo compartilhar de forma interativa este acervo de informações geoespaciais. A seguir, serão descritas as informações temáticas do BDiA.

4.1.1 Geologia

Existem inúmeras informações geológicas no acervo do BDiA. As informações temáticas de Geologia disponíveis e utilizadas na pesquisa foram: Unidades de

Mapeamento/Área de Geologia e tempo geológico. As Unidades de Mapeamento de Geologia estão consolidadas em um dado geoespacial vetorial representado por feições de polígonos. Seu conteúdo consiste no mapeamento geológico contínuo organizado em um sistema de classificação hierarquizado em três níveis taxonômicos que considera identificação, distribuição e escala de análise das unidades de mapeamento. Na Figura 4.2 está demonstrado o nível 1, ou o 1° táxon, que corresponde Substrato Geológico do Cerrado.

Figura 4.2 – Substrato Geológico do Bioma Cerrado.



Fonte: IBGE (2023).

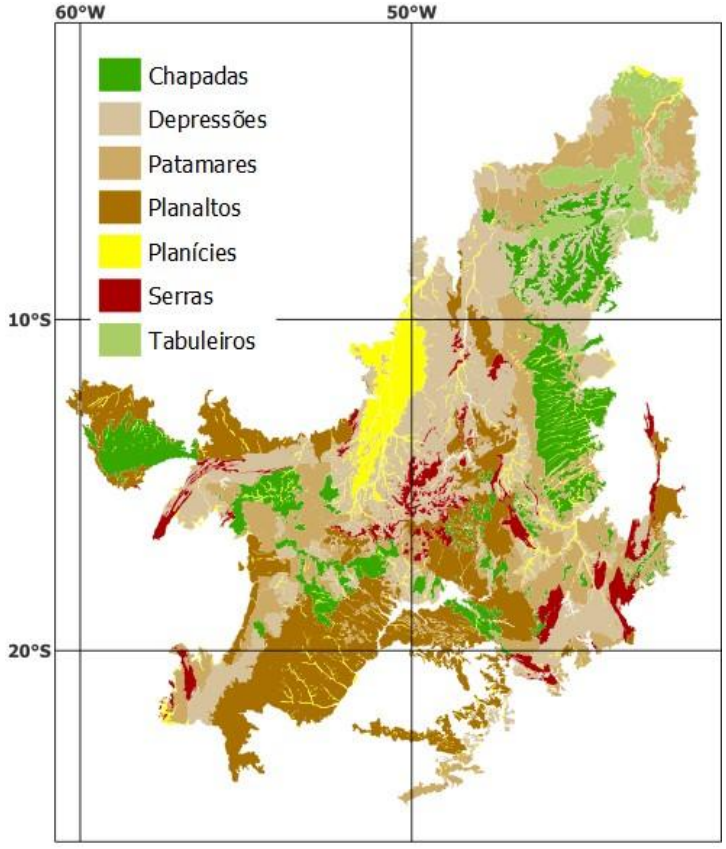
4.1.2 Geomorfologia

As Unidades de Mapeamento de Geomorfologia estão consolidadas em um dado geoespacial vetorial representado por feições de polígonos. Seu conteúdo consiste no mapeamento geomorfológico contínuo organizado em um sistema

de classificação hierarquizado em quatro níveis taxonômicos: o primeiro nível taxonômico corresponde aos Domínios Morfoestruturais; o segundo nível taxonômico corresponde às Regiões Geomorfológicas; o terceiro nível taxonômico corresponde às Unidades Geomorfológicas; e o quarto nível taxonômico e menor unidade de mapeamento de geomorfologia corresponde aos Modelados de Relevo. Estão contemplados atributos básicos como símbolo, categoria, natureza, característica, forma, altimetria e dissecação do modelado.

A Figura 4.3 mostra as informações da Geomorfologia, referentes ao compartimento de relevo, do Bioma Cerrado.

Figura 4.3 – Compartimentação de relevo do Bioma Cerrado.



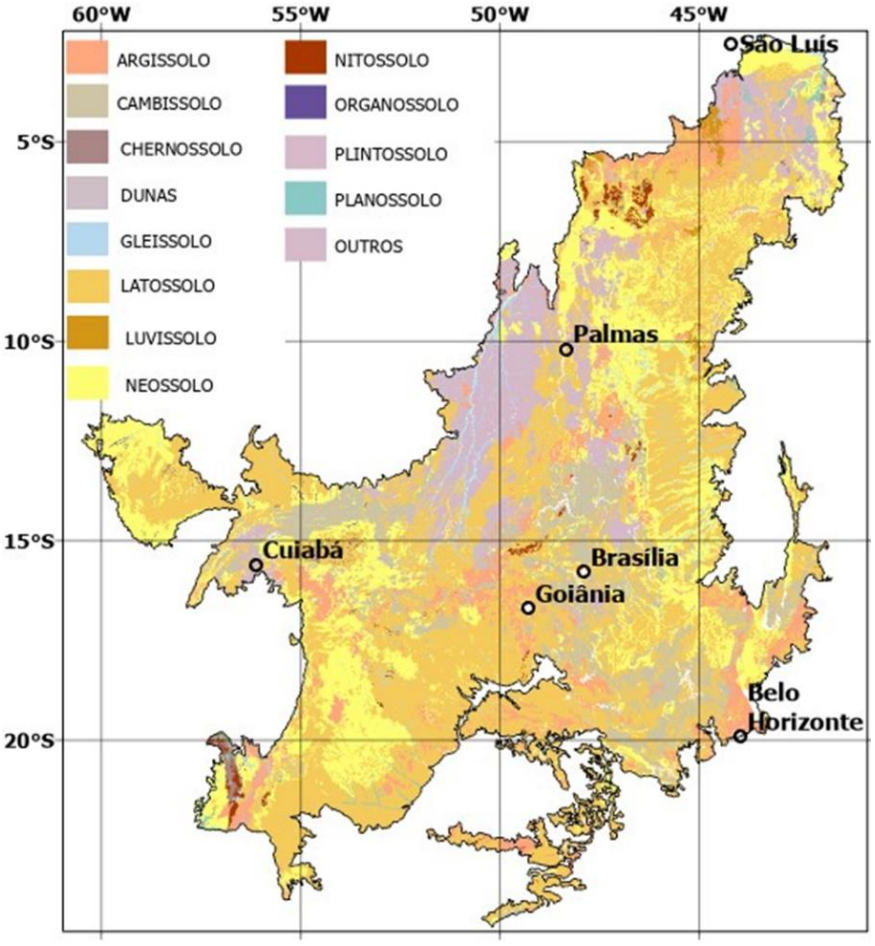
Fonte: IBGE (2023).

4.1.3 Solos

As Unidades de Mapeamento de Pedologia estão consolidadas em um dado geoespacial vetorial representado por feições de polígonos. Seu conteúdo

consiste no mapeamento pedológico contínuo organizado em um sistema de classificação hierarquizado em quatro níveis categóricos que considera seus atributos, horizontes diagnósticos e suas características morfológicas, físicas e químicas, seguindo a classificação taxonômica do SiBCS (Sistema brasileiro de classificação de solos). O primeiro nível categórico corresponde às Ordens; o segundo nível categórico corresponde às Subordens; o terceiro nível categórico corresponde aos Grandes Grupos; o quarto nível categórico corresponde aos Subgrupos. A Figura 4.4 mostra as classes de Ordem de solos do Bioma Cerrado.

Figura 4.4 – Solos do Bioma Cerrado.

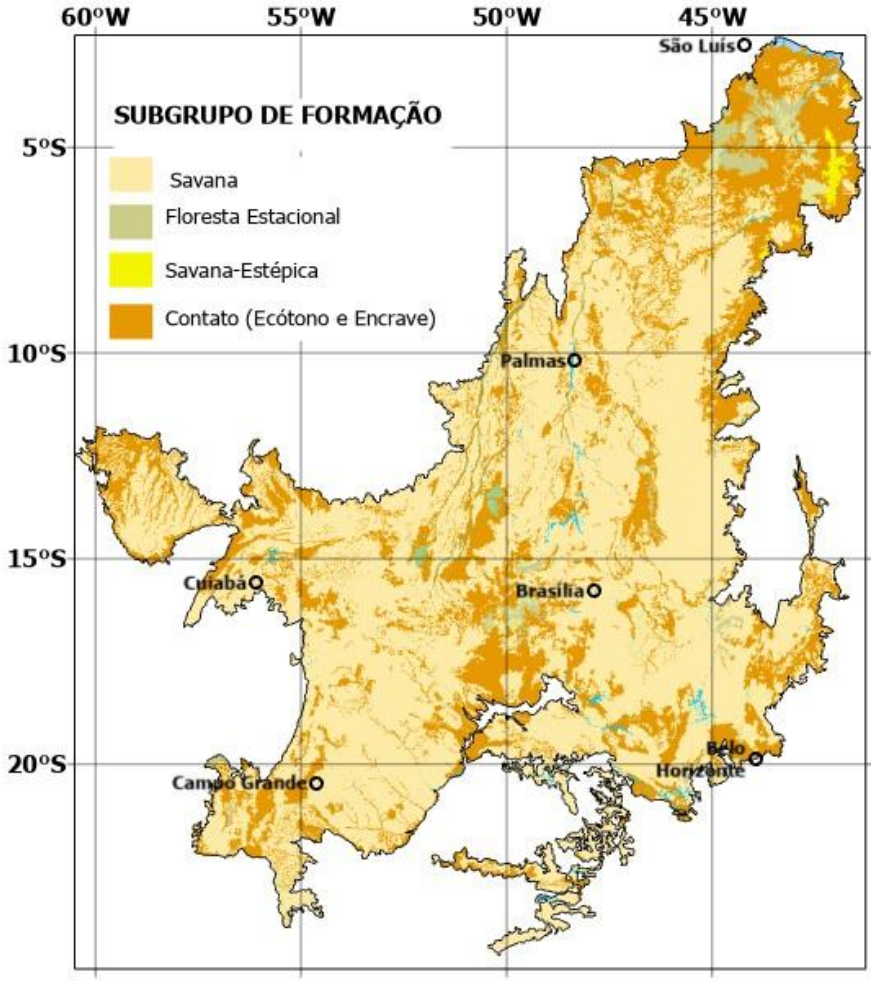


Fonte: IBGE (2023).

4.1.4 Vegetação

As Unidades de Mapeamento de Vegetação, no BDiA, estão contempladas em atributos como: legenda (composta por símbolos de um ou mais classes de legenda da SCVB, sejam naturais ou antrópicos); tipo de classe de cobertura e uso predominante (vegetação natural ou área antrópica predominante com ou sem presença de tensão ecológica); Regiões Fitoecológicas e outras áreas; 1ª, 2ª e 3ª Cobertura vegetal ou outras áreas (predominante); classe de vegetação ou área natural (quando predominante); classe de área antrópica (quando predominante); classe de contato (quando existente); vegetação pretérita (até o nível identificável); entre outros. A Figura 4.5 mostra um dos níveis das unidades de mapeamento, o Subgrupo de Formação da Vegetação do Bioma Cerrado.

Figura 4.5 – Subgrupo de Formação de vegetação do Bioma Cerrado.



Fonte: IBGE (2023).

No portal BDiAWeb do tema Vegetação, as camadas de informação do mapeamento podem ser visualizadas em 3 níveis: 1° Nível (escala de visualização na tela menor que 1: 5.000.000): Regiões Fitoecológicas e outras áreas. As Regiões Fitoecológicas mapeadas no território nacional são: Floresta Ombrófila Densa, Floresta Ombrófila Aberta, Floresta Ombrófila Mista, Floresta Estacional Sempre-Verde, Floresta Estacional Semidecidual, Floresta Estacional Decidual, Campinarana, Savana, Savana-Estépica e Estepe. São também representadas as outras áreas, relevantes no mapeamento e que não se enquadram no conceito de Região Fitoecológica. Essas áreas são as Formações Pioneiras e as áreas de Contatos entre diferentes tipos de vegetação (Ecótonos e Encraves).

2° Nível (escala de visualização na tela entre 1: 5.000.000 e 1:400.000): Cobertura Vegetal e outras áreas (predominante). Para a vegetação natural, essa simplificação corresponde ao nível hierárquico de **formação** na Classificação da Vegetação Brasileira do IBGE.

3° Nível (escala de visualização na tela maior que 1:400 000): Cobertura Vegetal e outras áreas. Corresponde à classificação da Cobertura Vegetal e outras áreas, em seu maior detalhamento para a escala 1:250 000.
O tema de vegetação no BDiA está baseado nos levantamentos de campo realizados pelo IBGE e descritos no Manual Técnico da Vegetação Brasileira do IBGE (IBGE, 2012).

4.1.4.1 As unidades de Savana no BDiA

A Savana (Cerrado) é conceituada como uma vegetação xeromorfa, que ocorre sob distintos tipos de clima. Reveste solos pobres em nutrientes e ricos em alumínio, com comunidades de plantas adaptadas, como gramíneas, plantas rasteiras, arbustos pequenos e vegetação com estruturas subterrâneas. Essas plantas, geralmente de pequeno porte e adaptadas a condições difíceis, ocorrem em toda a região Neotropical, com maior predominância no Brasil Central. Em outras partes do País, recebe nomes locais, como: “Tabuleiro”, “Agreste” e “Chapada”, na Região Nordeste; “Campina” ou “Gerais” no norte dos Estados de Minas Gerais, Tocantins e Bahia; e “Lavrado” no Estado de Roraima, entre outras

denominações (IBGE, 2012). No BDiA a Savana (Cerrado) foi subdividida em quatro subgrupos de formação: Florestada; Arborizada; Parque; e Gramíneo-Lenhosa (Figura 4.6).

4.1.4.1.1 A Savana florestada

Encontra-se em solos profundos e pobres em nutrientes, típicos de climas tropicais com estações bem definidas. Contém comunidades de plantas lenhosas pequenas, com ramificação irregular e casca grossa (suberizada) ou ligeiramente macia. Inclui árvores ou arbustos com alturas que variam entre 6 e 8 metros, além de plantas que armazenam nutrientes em estruturas subterrâneas. Em algumas áreas, há vegetação lenhosa com altura superior a 10 metros, visualmente semelhante a florestas estacionais, mas com composição diferente. Não há uma camada nítida de plantas rasteiras, mas sim vegetação composta por gramíneas e pequenos arbustos, além de palmeiras anãs espalhada.

4.1.4.1.2 Savana arborizada

Trata-se de um subgrupo de vegetação natural ou antropizado, com uma composição que combina árvores e arbustos pequenos e esparsos, além de um estrato de aparência graminosa de forma contínua. Essa vegetação é frequentemente sujeita a queimadas anuais. As comunidades predominantes ora podem formar paisagens mais abertas, conhecidas como Campo Cerrado, ora áreas com vegetação mais densa, chamadas Cerrado propriamente dito.

4.1.4.1.3 Savana parque

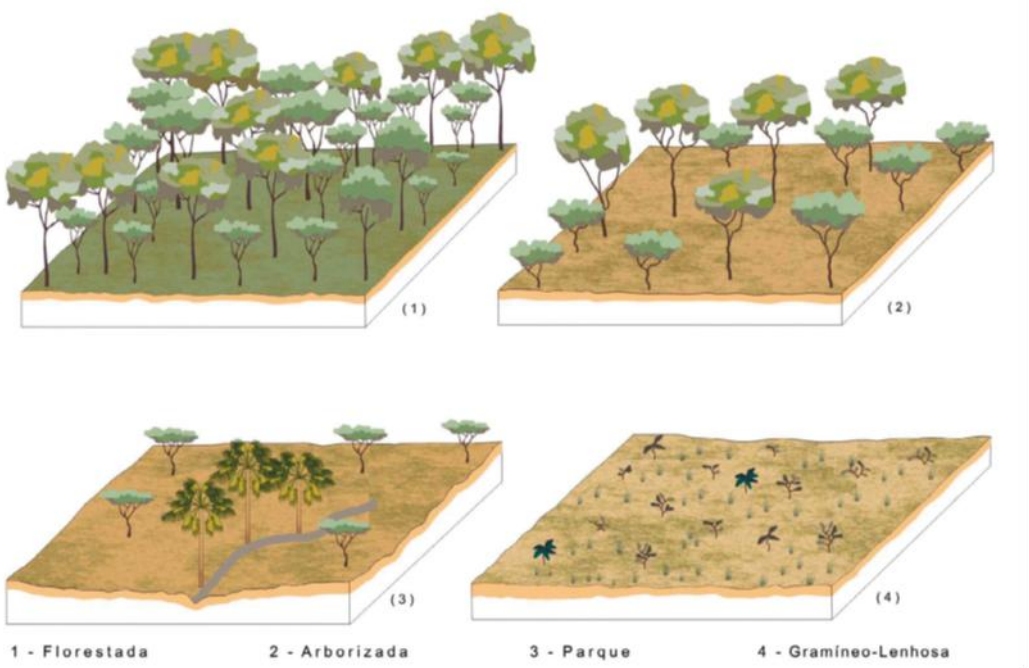
Subgrupo de formação constituído essencialmente por um estrato de aparência graminosa, integrado por vegetação rasteira e vegetação com tubérculos e raízes profundas de florística natural ou antropizada, entremeado por vegetação lenhosa de pequeno porte isolada, com conotação típica de um “Parque Inglês” (*Parkland*). A Savana Parque de natureza antrópica é encontrada em todo o País, enquanto a natural ocorre algumas, vezes, com feição de campos litossólicos e/ou rupestres. Em áreas encharcadas de depressões periodicamente inundadas, ocorrem as tipologias naturais de Cerrado-de-

Pantanal, com denominações regionais diversas, caracterizadas pela presença de “covoais”, “monchões” ou “murundus”.

4.1.4.1.4 Savana gramíneo-lenhosa (Campo-Limpo-de-Cerrado)

Quando natural, nesta fisionomia predominam os gramados entremeados por plantas lenhosas raquíticas, que ocupam extensas áreas dominadas por vegetação e que, aos poucos, quando manejados através do fogo ou pastoreio, vão sendo substituídos por vegetações que se distinguem por apresentar colmos subterrâneos, portanto mais resistentes ao pisoteio do gado e ao fogo.

Figura 4.6 - Blocos-Diagramas das fisionomias ecológicas da Savana, segundo o IBGE.



Fonte: IBGE (2012).

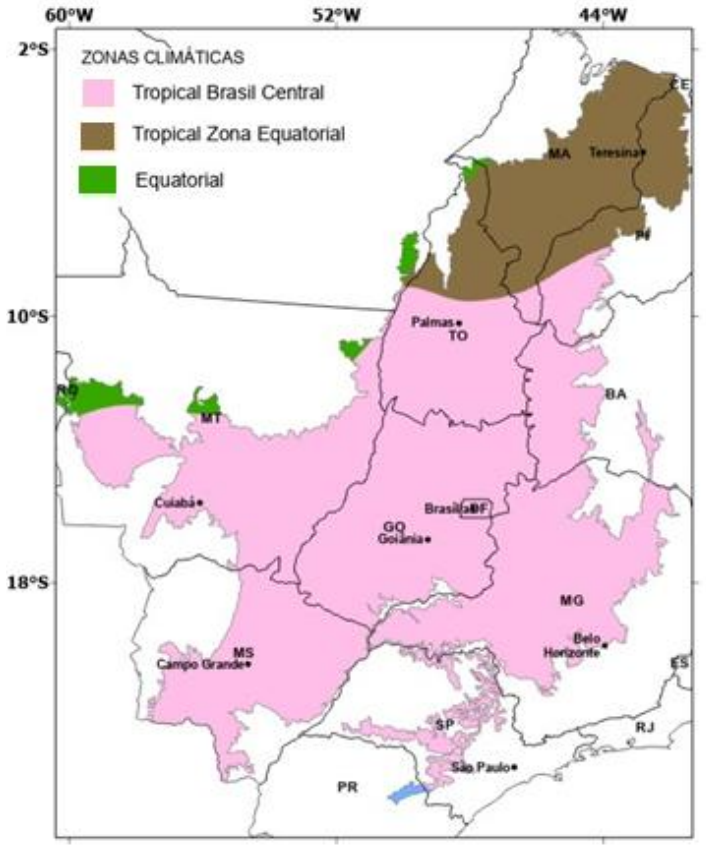
4.2 Informações ambientais de Clima - IBGE

Entre as informações ambientais fundamentais para o estudo do Cerrado, está o estudo do clima. Para efeito desta pesquisa, para manter a coerência do levantamento dos recursos naturais que compõe a pesquisa com dados advindos do BDiA, optou-se pela escolha dos dados de clima, em escala 1:5.000.000, que estão disponíveis no site do IBGE no mapa de clima da

Diretoria de Geociências do IBGE, lançado em 2002. O mapa elaborado pelo Departamento de Recursos Naturais e Estudos Ambientais (à época), entretanto, foi uma adaptação do Mapa Brasil Climas do IBGE, de 1978, utilizando a classificação de Nimer (1979). Infelizmente, não há muita informação dos metadados desta documentação cartográfica, além do que está descrito na legenda do mapa.

A Figura 4.7 mostra a classificação climatológica de Nimer (1979) para o Bioma Cerrado.

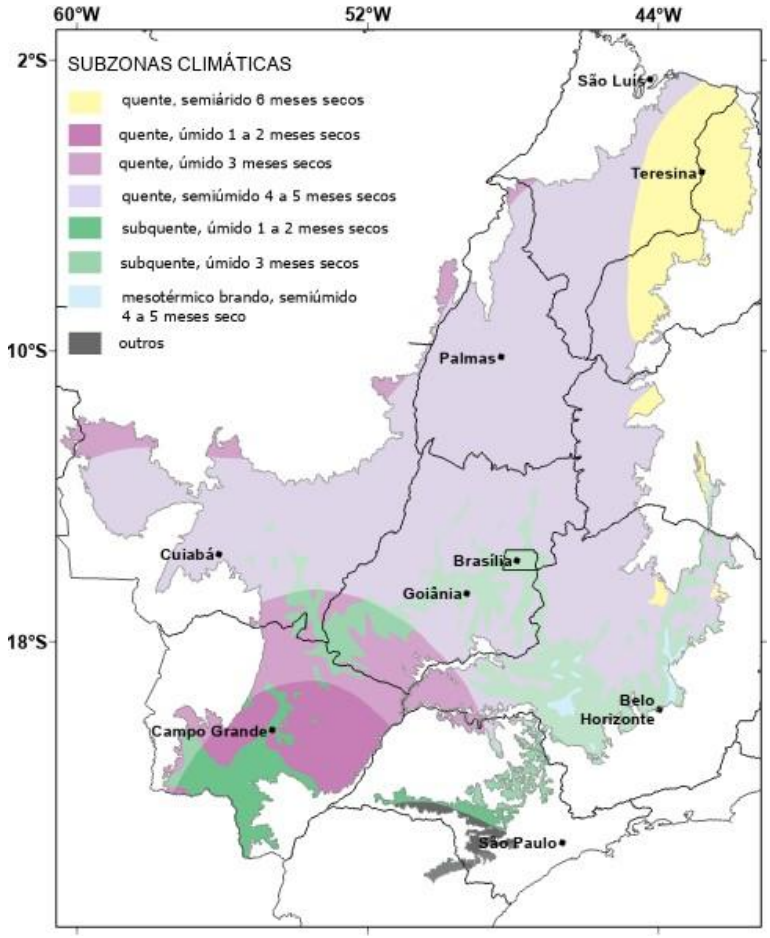
Figura 4.7 - Zonas climáticas para o Cerrado, segundo Nimer (1979).



Fonte: IBGE (2024).

A Figura 4.8 mostra a classificação climática de Nimer para o Bioma Cerrado, em função das subzonas climáticas, onde são levadas em consideração a temperatura média e a quantidade de meses secos durante o ano, em função da precipitação.

Figura 4.8 – Subzonas climáticas para o Bioma Cerrado, segundo Nimer (1979).



Fonte: IBGE (2023).

A classificação climática de Nimer (1979) propõe uma metodologia específica para categorizar os climas no Brasil, levando em consideração aspectos como temperatura, precipitação e a distribuição sazonal das chuvas. No contexto do bioma Cerrado, essa classificação oferece uma perspectiva detalhada sobre as variações climáticas da região.

Segundo Nimer (1979), o Cerrado é predominantemente caracterizado por um clima tropical com estação seca bem definida. A metodologia de Nimer, para determinar os meses secos, considera como seco o mês em que a precipitação é menor que o dobro da temperatura média desse mês. Essa abordagem difere de outras classificações, como a de *Köppen*, que classifica grande parte do Cerrado como clima tropical de savana com inverno seco (Aw), caracterizado

por temperaturas médias acima de 18°C em todos os meses e precipitação no mês mais seco inferior a 60 mm.

Apesar da classificação climática de Nimer (1979) apresentar limitações e críticas associadas à sua metodologia, quando aplicada a estudos ambientais e de planejamento territorial, ela oferece uma visão detalhada sobre as variações climáticas do Cerrado.

4.3 Grade estatística de informações ambientais - IBGE

As grades estatísticas constituem-se em uma forma de espacialização, armazenamento e recuperação de dados em pequenas unidades geométricas – no caso quadrados – que representam recortes específicos do mundo real, e que funcionam como blocos de construção; dispõem de uma inerente estabilidade espaço-temporal, e são, enquanto matrizes, extremamente adequadas à conexão a bancos de dados. Por sua natureza modular, podem representar fenômenos ou propriedades do terreno, independentemente dos limites político-administrativos ou operacionais em que ocorram (município, distrito, jurisdição, setor, etc.).

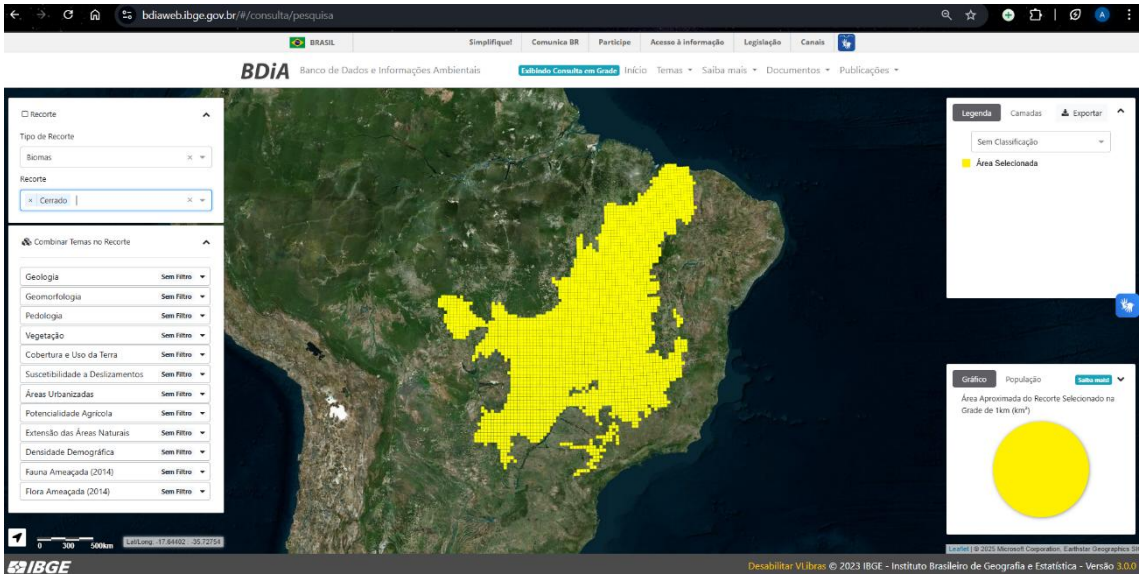
O produto Grade Estatística utilizado como fonte (IBGE, 2016) é disponibilizado em grades divididas por quadrantes de 500 km x 500 km; cada quadrante é composto por células quadradas de 1 km² (1 km x 1 km) em áreas rurais e 0,04 km² (200 m x 200 m) em áreas urbanas; sua tabela de atributos possui campos que referenciam cada célula ao identificador do respectivo superior hierárquico, permitindo a agregação das referidas células, em células de 5 km x 5 km, 10 km x 10 km, 50 km x 50 km, 100 km x 100 km e 500 km x 500 km. Este dado foi elaborado originalmente em um sistema de coordenadas métricas e na projeção cartográfica cônica equivalente de Albers, de modo a preservar a área de cada célula, independentemente da posição geográfica, sendo posteriormente reprojetada para o sistema de coordenadas geográficas com projeção cilíndrica equirretangular (IBGE, 2016; 2023).

A Grade Estatística de Dados Ambientais é um módulo ainda experimental da plataforma BDiA. Sua estrutura viabiliza a integração de informações ambientais em Grade Estatística, que abrange diferentes insumos ambientais e recortes

territoriais. O conjunto de dados geoespaciais está distribuído em pastas por resolução espacial de células da grade em estrutura vetorial no formato Shapefile. Os dados estão disponibilizados no recorte territorial do Brasil, em resoluções de 25 km a 1 km, e as resoluções de 500 m e 250 m estão disponibilizados por Unidades da Federação, a fim de viabilizar, além da disseminação do dado, o seu carregamento e visualização em ambiente SIG, devido a quantidade de células nas referidas resoluções (IBGE, 2023).

A versão do BDiA (*release* de novembro de 2023) já permite ao usuário fazer consultas e fazer o *download* dos dados em formato de grade (Figura 4.9).

Figura 4.9 - Visão da Grade Estatística de Dados Ambientais, do IBGE, com o recorte do Bioma Cerrado.



Fonte: Elaborado pela autora.

4.4 Informações ambientais da EMBRAPA

4.4.1 Água disponível no solo

O Cerrado, frequentemente chamado de a "savana brasileira", está associado a desafios climáticos e edáficos únicos que tornam o estudo da água disponível no solo essencial para sua sustentabilidade e para as atividades humanas associadas a ele.

A água disponível no solo no Cerrado desempenha um papel crucial na manutenção de seus ecossistemas e na produtividade agrícola da região. Este bioma é caracterizado por períodos de seca prolongada, alternados com chuvas intensas concentradas em poucos meses do ano, o que exige que a vegetação nativa desenvolva estratégias específicas para lidar com a limitação hídrica.

4.4.2 Definição das classes de Água Disponível (AD) no solo

A água disponível no solo (AD) é a água que os solos podem armazenar no seu espaço poroso de modo que as plantas possam acessá-la, garantindo a vida continental no planeta.

Segundo Araújo Filho et al. (2022), a AD é um parâmetro de grande importância agronômica, ecológica, meteorológica, hidrológica, bem como para predição de safras. Destaca-se também, como um parâmetro relevante para estudos de erosão do solo, cobertura vegetal, proteção ambiental, estoque de carbono e mudanças climáticas. Além disso, é um importante indicador de qualidade física de solos.

A classificação de AD, elaborada pela EMBRAPA SOLOS, está categorizadas em 7 classes, e é baseada em faixas de valores expressos em mm de água por cm de solo. Segundo o estudo da Embrapa, são utilizadas as seguintes classes para mapeamento (Figura 4.10):

AD0 (< 0,34 mm/cm) - **Extremamente Baixa**.

Características: Solos muito arenosos ou rasos, com baixa retenção de água. A infiltração da água ocorre rapidamente, mas a capacidade de armazenamento é mínima. Exemplos: Neossolos Quartzarênicos e algumas classes de Areias Quartzosas.

AD1 (0,34 – 0,47 mm/cm) – **Baixa**.

Características: Solos predominantemente arenosos, mas com alguma capacidade de retenção hídrica devido à presença de frações finas (silte e argila). Exemplos: Neossolos Flúvicos arenosos e algumas variantes de Latossolos pouco desenvolvidos.

AD2 (0,48 – 0,67 mm/cm) - **Moderadamente Baixa**.

Características: Solos de textura média a arenosa-franca, com maior equilíbrio entre drenagem e retenção de água. Exemplos: Cambissolos arenosos e alguns Argissolos com baixa presença de argila.

AD3 (0,68 – 0,94 mm/cm) – **Moderada**.

Características: Solos franco-arenosos ou franco-argilosos, oferecendo boa retenção de água sem comprometer a drenagem. Exemplos: Argissolos e Latossolos Amarelos de textura média.

AD4 (0,95 – 1,32 mm/cm) - **Moderadamente Alta**.

Características: Solos bem estruturados, com alta capacidade de retenção de água e boa disponibilidade para as plantas. Exemplos: Latossolos Vermelhos e Argissolos bem estruturados.

AD5 (1,33 – 1,84 mm/cm) – **Alta**.

Características: Solos argilosos e siltosos, com grande capacidade de retenção de água, mantendo umidade por longos períodos. Exemplos: Nitossolos, Gleissolos e algumas variantes de Latossolos muito argilosos.

AD6 (> 1,84 mm/cm) - **Extremamente Alta**.

Características: Solos de textura muito argilosa, com alta retenção de umidade. O excesso de água pode dificultar a aeração das raízes. Exemplos: Vertissolos e Gleissolos mal drenados.

Figura 4.10 – Padrão de cores e intervalos das classes de Água Disponível no solo.

Classe	
Símbolo / cor	Intervalo (mm/cm)
AD0	< 0,34
AD1	0,34 a 0,47
AD2	0,48 a 0,67
AD3	0,68 a 0,94
AD4	0,95 a 1,32
AD5	1,33 a 1,84
AD6	> 1,84

Fonte: Adaptado de Araújo-Filho et al. (2022).

Nesta pesquisa, optou-se pelas informações de água disponível no solo, de 100 a 200 cm, do repositório da EMBRAPA SOLOS por dois motivos: pela compatibilidade da escala de 1:250.000, ou seja, a mesma do BDiA, e pela disponibilidade das informações serem elaboradas em classes, o que facilita a transposição para dados vetoriais (categóricos). O projeto original da EMBRAPA disponibilizou, em resolução espacial (pixel) de 90m, as classes de AD para as profundidades de solo de 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm e 100-200 cm.

O critério de escolha da profundidade de água disponível no solo foi feito em função das raízes da vegetação do cerrado serem profundas, em busca de água, e pela compatibilidade temporal com os demais dados, uma vez que os temas, em sua maioria, são de levantamentos dos anos 70 e 80. Assim, quanto mais profundos os dados de água disponível, menor a probabilidade de armazenamento de água no solo ter sofrido intervenção antrópica, tornando mais robusta a análise de disponibilidade hídrica em relação à época dos levantamentos dos demais dados temáticos.

4.5 Informações ambientais do CPRM/SGB

A declividade do terreno é um fator essencial na caracterização de paisagens naturais e agrícolas, influenciando processos como drenagem, erosão e aptidão agrícola. A Embrapa classifica a declividade do solo em categorias padronizadas, permitindo sua aplicação em planejamento territorial, conservação ambiental e manejo agrícola.

4.5.1 Classificação de declividade da EMBRAPA

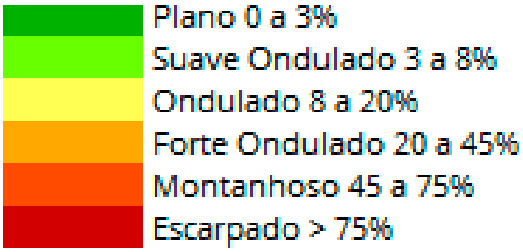
As classes de declividade seguem uma divisão baseada em percentuais de inclinação, categorizadas pela EMBRAPA (1979) da seguinte forma:

- Plano a suavemente ondulado (0 - 3%)
- Ondulado (3 - 8%)
- Forte ondulado (8 - 20%)
- Montanhoso (20 - 45%)
- Escarpado (> 45%)

Essa classificação é amplamente utilizada para zoneamento agrícola e análise de risco de erosão. Na pesquisa, foi feita a escolha destas classes de declividade em percentagem, por facilitar a transposição para classes vetoriais e organização do Banco de Dados Geográficos.

A Figura 4.11 mostra a padronização de cores por classe de declividade, conforme EMBRAPA (1979).

Figura 4.11 – Classes de Declividade em percentagem e suas cores.

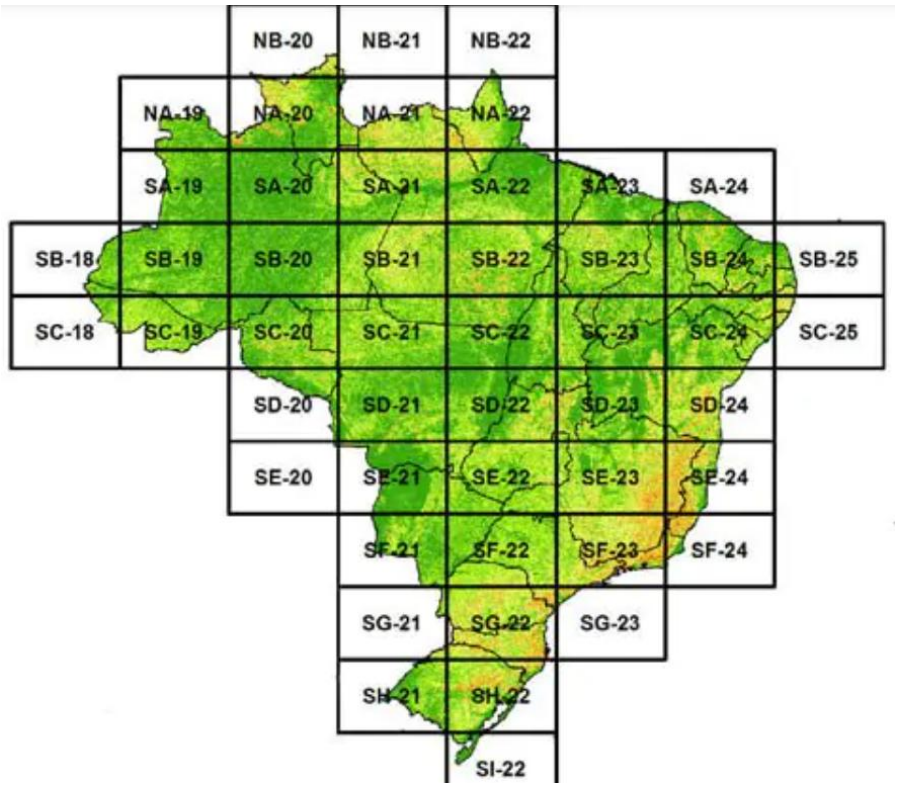


Fonte: Adaptado de EMBRAPA (1979).

4.5.2 Dados de declividade da CPRM

A CPRM (Serviço Geológico do Brasil) disponibiliza mapas de declividade (SRTM 90m) no portal da SGB (https://www.sgb.gov.br/mapa-de-declividade-em-percentual-do-relevo-brasileiro), onde é possível acessar dados em formato raster, em cartas ao milionésimo (Figura 4.12).

Figura 4.12 – Cartas de Declividade por articulação ao Milionésimo.



Fonte: CPRM (2023).

As informações dos mapas têm aplicação em estudos de aptidão agrícola, riscos geotécnicos e conservação ambiental. Em especial, a declividade mostra-se substancial na implantação de bases de dados para tomada de decisão no desenvolvimento de um Sistema Nacional de Monitoramento e Alerta de Desastres Naturais, e também um importante instrumento de análise ambiental para o estudo do espaço geográfico de modo integrado.

A declividade do terreno é um elemento topográfico importante para a distribuição das fitofisionomias do Cerrado, pois influi diretamente na dinâmica hídrica, a erosão do solo e a disponibilidade de nutrientes. Como o Cerrado apresenta grande heterogeneidade ambiental, variações no gradiente de declividade podem resultar em diferenças significativas na composição e estrutura da vegetação.

5 MATERIAIS E MÉTODOS

Neste capítulo serão abordados os materiais utilizados no experimento da pesquisa, que foi realizado em duas etapas: a primeira, direcionada para os dados da Reserva Ecológica do IBGE (RECOR), como área piloto, para o desenvolvimento da metodologia, e uma segunda etapa, com as informações e dados para os processamentos do Bioma Cerrado advindos do Banco de Dados Ambientais do IBGE, juntamente com dados de outras instituições responsáveis pela elaboração de dados de informações ambientais oficiais, EMBRAPA e CPRM.

A metodologia da pesquisa, em grande parte, foi uma metodologia "*ad hoc*", uma vez que foi desenvolvida com abordagens ajustadas especificamente para atender às questões e objetivos do trabalho, devido à sua natureza inédita. Entretanto, a fase inicial da pesquisa foi embasada em uma adaptação da ideia da metodologia proposta por Barbosa et al. (1998), no que diz respeito a integração de temas com técnicas de geoprocessamento e sensoriamento remoto.

A proposta metodológica de Barbosa et al. (1998), por sua vez, baseia-se nos princípios da metodologia de estudos integrados de levantamentos de recursos naturais, que é um dos carros-chefe do IBGE desde os primórdios dos levantamentos de campo, na década de 1970. Na integração de temas cartografados pelo IBGE, realizada em levantamentos de campo, há uma sequência sistemática de integração conduzida pelos especialistas temáticos, tanto durante os trabalhos “in loco” quanto no laboratório.

Nesta sistematização, a geologia é considerada a "ciência-mãe" e serve como base para os demais temas. Ou seja, as informações geológicas orientam a integração de temas como geomorfologia, pedologia e, finalmente, vegetação. Essa sequência é preservada na metodologia de integração de temas de Barbosa et al. (1998), utilizando como elemento de verdade terrestre as imagens de satélite, que permitem a delimitação da Unidade Territorial Básica (UTB).

A metodologia de Barbosa et al. (1998) foi amplamente empregada nos Zoneamentos Ecológicos-Econômicos (ZEE) e nos levantamentos da RIDE-DF,

realizados pelo Ministério do Meio Ambiente (MMA) nas décadas de 1990 e 2000.

A pesquisa foi dividida em dois momentos, onde na primeira parte foram realizados os processamentos utilizando-se dos dados temáticos da RECOR[3] e, em um segundo momento, os processamentos com os dados de Savana do Bioma Cerrado, do BDiA. Neste capítulo, a sequência descrita apresenta na Seção 5.1, uma breve descrição da área da RECOR (e entorno) e a justificativa de sua escolha. Na Seção 5.2 são discorridas as etapas metodológicas elaboradas durante a pesquisa, com os dados da RECOR, sobre as regras de associação, Árvore de Decisão e Random Forest no aplicativo WEKA. Finalmente, na Seção 5.3, constam as etapas metodológicas realizadas com os dados do bioma Cerrado.

5.1 Descrição da área da RECOR

A Reserva Ecológica do IBGE (RECOR) é uma área protegida de interesse científico sob a gestão do IBGE. Criada em 22 de dezembro de 1975, a Reserva tornou-se referência na produção de informações ambientais motivada pela preocupação em fornecer subsídios ao planejamento territorial sustentável para o bioma Cerrado. Desde 1998, integra o grupo das estações de pesquisas ecológicas de longa duração que compõem redes nacionais e continentais interligadas à rede ILTER (*Internacional Long Term Research Network*).

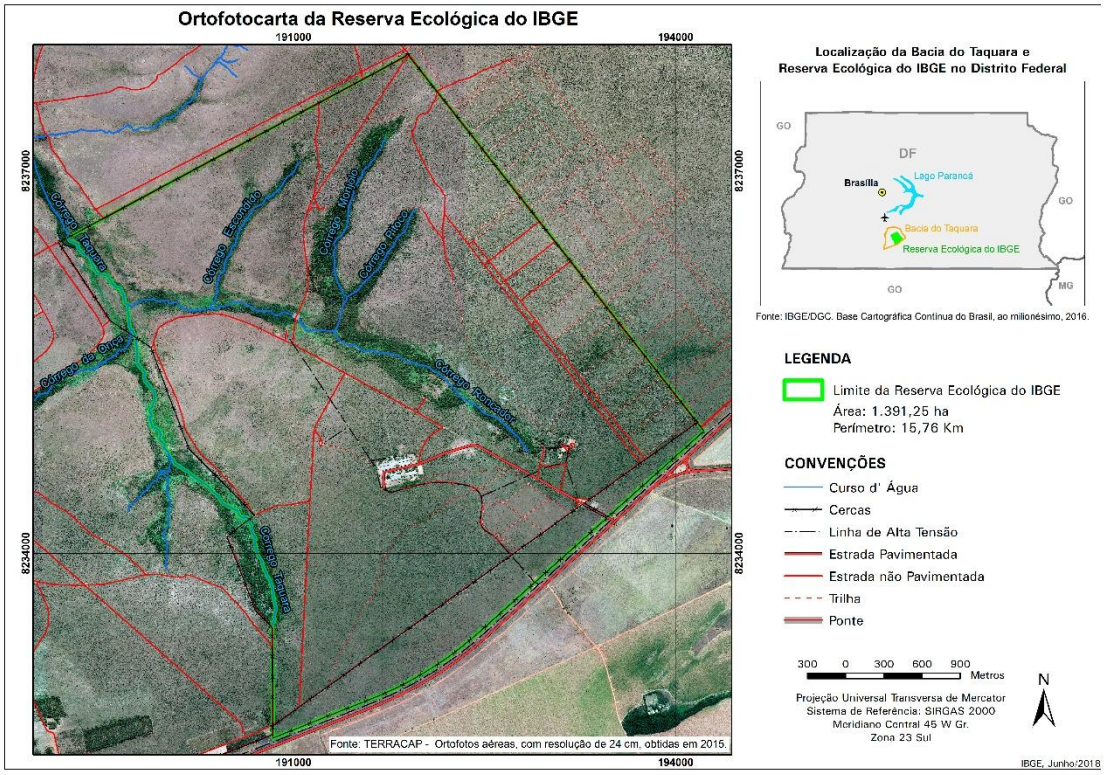
A RECOR está localizada a 26 km do centro de Brasília, com acesso pela Estrada Parque Contorno (DF-001). Situada na bacia do Lago Paranoá, a área foi originalmente criada como uma faixa de proteção para os mananciais hídricos da capital. A RECOR é drenada por cinco córregos, protegidos por Matas de Galeria, Veredas e uma matriz de Cerrados, que garantem a qualidade da água.

[3] Os dados digitais temáticos da RECOR, utilizados na pesquisa, foram solicitados à equipe do IBGE responsável pela Reserva. Este material abrange toda a Bacia Hidrográfica do córrego Taquara, no Distrito Federal, próximo à Brasília. Assim, a pesquisa trabalhou com dados ambientais em uma área de abrangência para além dos limites da RECOR.

A Reserva possui uma grande diversidade de fitofisionomias do bioma Cerrado, incluindo Cerrado Denso, Cerrado Típico, Cerrado Ralo, Campo Sujo, Campo Limpo, Brejos, Veredas e Matas de Galeria. Sua flora e fauna são altamente representativas do bioma, com uma riqueza de espécies comparável a regiões de alta biodiversidade, como a Amazônia, a Mata Atlântica e outras florestas tropicais do mundo (ex.: Barro Colorado no Panamá, La Selva na Costa Rica e Cocha Cashu no Peru).

A Figura 5.1 mostra a localização geográfica da Reserva, no Distrito Federal, em uma imagem que permite observar seu entorno em área de cerrado.

Figura 5.1 – Ortocarta de localização geográfica da RECOR.



Fonte: IBGE (2025).

5.2 A etapa experimental com os dados da RECOR

O experimento da RECOR foi dividido em três fases. A primeira fase foi usada para gerar o conjunto de dados transacionais, que forma a base para o experimento. Esta fase inclui três subfases: a elaboração do Banco de Dados

Geográficos, no QGIS 3.36, onde foram baixadas as informações temáticas da RECOR (limite, drenagem, curvas de nível, geologia, geomorfologia, pedologia e vegetação) e geradas a declividade, sendo criado mais um tema, o tema declividade. Uma segunda subfase, a geração das grades e integração dos temas. A terceira, que foi a exportação da tabela de atributos, em arquivos csv, para serem transformados em arquivos formato “arff” para serem trabalhados no WEKA.

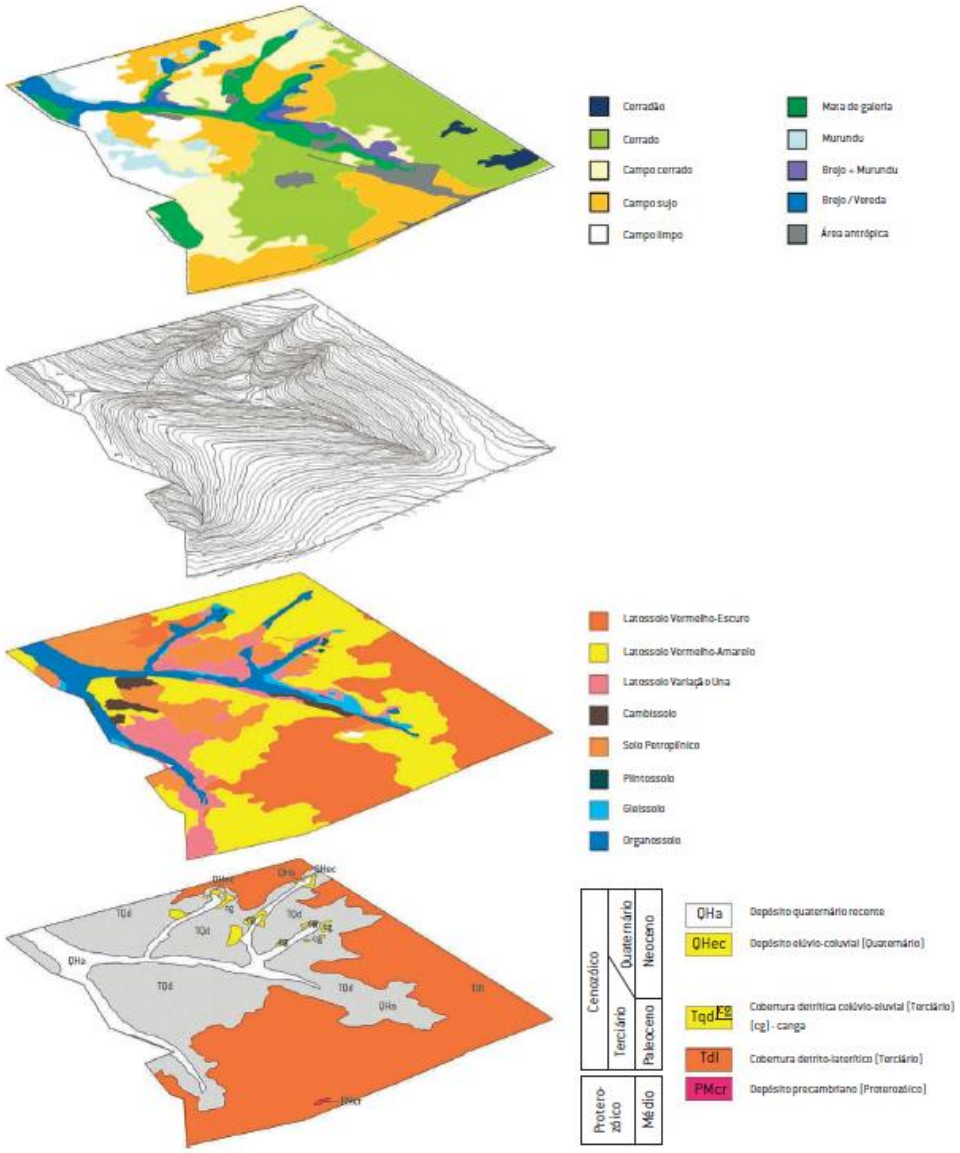
Na elaboração do Banco de Dados Geográficos, através da ferramenta do QGIS, foi gerada a camada de declividade, através das curvas de nível, sendo criadas quatro classes de declividade, vetoriais, de acordo com a Classificação da EMBRAPA (EMBRAPA, 1979) para declividade em percentagem. Em seguida, foi gerada a grade com agrupamento espacial de células de 20m x 20m. A grade foi processada com a ferramenta “gerar grades” do aplicativo QGIS. O tamanho das células da grade foi calculado para 20m x 20m, atendendo o critério PEC (Precisão de Exatidão Cartográfica) para a escala de mapeamento dos temas da RECOR, que foi elaborada em 1:10.000.

Para concluir a elaboração do conjunto de dados transacionais para as regras de associação, os dados a serem utilizados pelo algoritmo **Apriori**, no WEKA, precisavam ser **nominais** (categóricos). Assim, na integração dos temas foi utilizada a ferramenta “Associar atributos por localização” do QGIS. Este recurso processa uma camada vetorial de entrada e gera uma nova camada vetorial estendida. A camada resultante contém atributos suplementares incorporados à sua tabela de atributos, cujos valores são derivados de uma segunda camada vetorial. Um critério espacial é aplicado para selecionar os valores da segunda camada que são adicionados a cada recurso da primeira camada na nova camada resultante. No experimento, o critério espacial utilizado foi o de “maioria zonal”. A ferramenta permitiu que cada tema fosse integrado, em células, com dados nominais, pela sua maior área dentro da célula, formando nosso conjunto de dados transacionais.

Na nova grade gerada ainda foram realizadas as etapas de pré-processamento e limpeza dos dados (metodologia KDD), sendo finalmente gerada a grade final

com 105.026 células, integradas com os temas: geologia, geomorfologia, pedologia e declividade (Figura 5.2).

Figura 5.2 – Integração de temas da RECOR para o SIG.



Fonte: Adaptado de IBGE (2011).

Em seguida à limpeza dos dados, foi feita a exportação dos temas por células de grade, em arquivos csv, para o aplicativo WEKA. Cada célula passou a ser uma instância na tabela de atributos. No total, o conjunto de dados foi composto por 105.026 instâncias (linhas da tabela), sendo quatro variáveis (colunas) preditoras e uma variável-alvo.

As Regras de Associação foram geradas na segunda fase pelo algoritmo *Apriori* do WEKA. A métrica de *conviction* foi utilizada para filtrar regras desinteressantes para as classes-alvo de interesse (classes dos tipos de vegetação do cerrado).

Para facilitar a execução (geração) das regras de associação, as classes de interesse foram selecionadas por subconjuntos (*subsets*), através do filtro “SubsetByExpression”, no WEKA. A cada classe gerada como *subset*, estas iam sendo configuradas como “consequente” utilizando o hiperparâmetro CAR.

O CAR no WEKA é o acrônimo de *Class Association Rules* (Regras de Associação de Classe). Este recurso permite gerar regras de associação em que um atributo específico (geralmente chamado de "classe") é sempre colocado no consequente das regras. Assim, ele é útil para tarefas onde o objetivo é prever ou analisar a ocorrência de uma categoria específica, que era a intenção de nosso experimento. No experimento, cada classe de cerrado da RECOR foi filtrada em subconjuntos (*subset*) para obter as regras de associação por classe. Através do plugin ARviewer (visualizador de regras de associação) do WEKA, foram gerados gráficos das melhores regras para as classes de vegetação do cerrado da RECOR (vide exemplos de visualização de regras geradas pelo WEKA, no Apêndice C).

Após algumas análises, a investigação do experimento passou para uma outra abordagem, que foi e execução de outros algoritmos de Aprendizado de Máquina. Por também gerar regras, o conjunto de dados foi processado em um algoritmo de Árvore de Decisão. Foi escolhido, no WEKA, a árvore do algoritmo J48. A avaliação de desempenho dos algoritmos se dava através dos relatórios de desempenho gerados no próprio WEKA. Os relatórios gerados ajudavam na avaliação e ajustes da configuração dos hiperparâmetros, a fim de otimizar o desempenho do algoritmo, a cada execução, para alcançar resultados cada vez mais satisfatórios nas métricas de análise. Após algumas avaliações, o mesmo conjunto de dados foi submetido para execução do algoritmo Random Forest, também no WEKA. A cada execução, no aplicativo, iam sendo gerados os

relatórios de desempenho e de matriz de confusão, cujos resultados serão discutidos no Capítulo 6, na Seção 6.1.

Essas etapas de pré-processamento e de testes interpretativos são contempladas na Metodologia KDD, onde o conhecimento é construído com *feedbacks* em cada etapa. Segundo a metodologia KDD, é necessária a interpretação do que foi descoberto nos padrões, removendo redundâncias ou padrões irrelevantes. Deve-se buscar padrões vantajosos, úteis e compreensíveis para o usuário a fim de atingir o conhecimento a ser descoberto.

O experimento com os dados do Bioma Cerrado, praticamente, seguiu as mesmas etapas do experimento com os dados da RECOR. Entretanto, alguns processamentos foram necessários devido ao volume de dados a serem tratados.

5.3 A etapa experimental com os dados do Bioma Cerrado

O experimento do Bioma Cerrado foi dividido em três fases. A primeira fase foi usada para gerar o conjunto de dados transacionais, como na etapa anterior com os dados da RECOR. Esta fase incluiu três subfases: a elaboração do Banco de Dados Geográficos, no QGIS, onde foram baixadas as informações temáticas (geologia, geomorfologia, pedologia e vegetação) do Bioma Cerrado a partir do portal do BDiA, com o recorte pelo bioma Cerrado. Outros temas disponíveis em outros repositórios de dados foram baixados, tanto do IBGE (limite das unidades de federação, drenagem, clima), como de outras instituições (declividade) do CPRM e água disponível no solo (AD), profundidade de 100-200 cm, do repositório de dados da EMBRAPA SOLOS.

O limite do bioma, apesar de também estar disponível nos repositórios do IBGE, foi baixado do portal TerraBrasilis, do INPE.

Diferentemente da etapa experimental anterior, a declividade não foi gerada, mas baixada do repositório da CPRM, por articulação de Folha CIM (Carta ao milionésimo). As Folhas CIM baixadas e utilizadas na pesquisa, foram: SF22, SG21, SB22, SB23, SF21, SA24, SB24, SD20, SC23, SE23, SA23, SD21, SC21, SD23, SD22, SF23, SD24, SG22, SC22, SE22, SE21 e SE24.

No QGIS foram feitos os mosaicos das 22 folhas de declividade, em formato *raster*, sendo criado mais um tema, o tema declividade, que foi convertido em formato vetorial. Uma segunda subfase foi a geração da camada que continha as grades estatísticas. Da mesma forma que os dados de declividade, as grades estatísticas do IBGE foram acessadas a partir do repositório do IBGE e, no QGIS, foi feito o mosaico das grades estatísticas que abrangiam a área do Cerrado a fim de executar a integração dos temas. Convém salientar, que somente após esta fase do experimento, o IBGE lançou o *release* (de novembro de 2023) onde os dados já podem ser baixados no recorte da grade, pelo bioma de estudo, diretamente pelo BDiaWeb. Desta forma, não é preciso mais baixar cada pequeno *tile* da grade estatística de recobrimento do bioma de interesse e gerar um mosaico de toda a malha (recortes) de grade. Os dados que constam no release podem ser baixados e foram discretizados e codificados por valores numéricos.

A terceira subfase do experiemento foi a mais desafiadora, porque se tratava da exportação dos dados em formato de tabela (csv) para o aplicativo WEKA. A exportação da tabela de atributos, em arquivos csv, para serem transformados em formato *arff* (formato proprietário do WEKA) a fim de realizar as análises de regras de associação, foi toda executada em dados categóricos.

Um outro desafio foi trabalhar com a exportação de uma área de quase todo o Bioma Cerrado (aproximadamente 1,9 milhões de km^2), onde as células também eram de 1 km^2. Com um volume imenso de células, precisou ser feita uma pré-seleção de variáveis para o cruzamento. A princípio, foram selecionados 32 atributos das informações baixadas. O cruzamento foi feito com a mesma ferramenta do QGIS, “Associar atributos por localização”, divididos em sete cruzamentos dos temas com a grade estatística: geologia, geomorfologia, pedologia, clima, declividade, AWC (*Available Water Capacity*), ou seja, as AD (Água Disponível) e a vegetação. O cruzamento foi extremamente demorado devido ao volume de quase 2 milhões de células (área total do Bioma Cerrado, em km2) com armazenamento das informações dos atributos em forma de dados categóricos. Cada grupo de tema para cruzamento com a grade estatística

demorou, em média, 3 dias de processamento, sendo que alguns temas chegaram a 5 dias de processamento.

A etapa seguinte foi o cruzamento de todos os sete grupos de temas, a fim de se tornarem uma tabela de atributos única. Para este processamento foi necessário utilizar biblioteca Python, sendo a execução do processamento feito através da nuvem do Google *Colaboratory*. Uma vez integrados todos os temas, foi feito o pré-processamento de limpeza: remoção de valores ausentes, dados inconsistentes e a redução de valores redundantes com uso da biblioteca Python. Uma observação a ser feita é que pelo KDD, normalmente a integração é feita depois da limpeza, mas, no nosso experimento, teria que ser feito imediatamente após a espacialização (integração) por células, dado o longo tempo de processamento para o cruzamento, não tinha como ser feito com a informação ainda em forma de polígonos vetoriais dos sete temas.

A etapa do KDD de redução dos dados foi feita com nova seleção de atributos relevantes, ou seja, houve uma redução dimensional para simplificar a análise, mantendo a essência da informação, passando de 32 para 19 atributos, posteriormente mais uma agregação de variáveis e foi mantido no final 17 atributos como variáveis preditoras, acrescida da variável-alvo de interesse. A variável-alvo de interesse é a variável que será a predita. No caso, a vegetação de cerrado e as tipologias de savanas.

A redução dos dados com valores significativos, também foi feita com uma consulta utilizando a expressão:

"leg_uveg" LIKE 'S%' AND "nm_sec1" LIKE '%Savana%' AND "leg_sup" = 'Vegetação Natural Dominante', que resultou em 481980 registros. Depois, ainda foram deletadas variáveis que apresentavam mais de 240000 registros nulos.

O arquivo final do Bioma Cerrado, em formato csv, para ser trabalhado como conjunto de dados transacionais no WEKA ficou com **479.648** instâncias (células), sem registros nulos, todos em valores categóricos.

No WEKA foram realizadas as associações e geradas as regras para cada classe de cerrado. Assim como foi executado com o conjunto de dados da RECOR, o conjunto de dados do Bioma Cerrado passou para os

processamentos dos classificadores de Árvore de Decisão (J48) e Random Forest, após a geração das Regras de Associação. No caso do Random Forest, no WEKA, foi feita amostragem, porque o aplicativo não suportava o volume de dados para iterar um número muito grande de árvores. Desta forma, com o uso das ferramentas Python, foram realizadas amostragens estratificadas para execução de algoritmos, a fim de diminuir o tempo de processamento e assegurar que a distribuição das categorias no conjunto de treinamento mantivesse a proporção de classes do conjunto original. A amostragem estratificada é um método de pesquisa que divide o conjunto de dados em subgrupos, ou estratos, para garantir que cada um deles seja representado na amostra. Com amostragem estratificada de 10%, foram executados também, em códigos Python, os mesmos algoritmos classificadores para ter uma saída de espacialização dos dados para o QGIS.

Através do Python também foram realizados gráficos para as análises de MCA (*Multiple Correspondence Analysis*), gráficos de Cramer's V para dar apoio ao estudo dos tipos de savanas e compreender a variabilidade dos dados e correlações.

Com as análises dos relatórios, passou-se para a análise do classificador com melhor desempenho. Posteriormente, foi usado o classificador de Redes Bayesianas, tanto no WEKA, quanto nos códigos Python. No caso deste algoritmo, foram geradas as tabelas de Distribuição de Probabilidade por tipo de savana e também as redes bayesianas por grupo de variáveis. Após analisar os relatórios de desempenho e de matriz de confusão de cada algoritmo em busca do melhor desempenho do classificador, optou-se finalmente por um classificador híbrido, entre o Random Forest e o BayesNet, através do WEKA. Os resultados de cada etapa serão discutidos na Seção 6.2.

5.4 As classes de variáveis preditoras em cada experimento

Algumas variáveis preditoras utilizadas no experimento foram selecionadas *ad-hoc* pela autora, em função da disponibilidade operacional de processamento e de sua importância baseada na literatura (Apêndice A e B).

5.4.1 Variáveis da RECOR

Para a RECOR, as variáveis e classes foram as mesmas classes feitas pelos especialistas temáticos, na execução do levantamento em escala de 1: 8.000:

Classes da variável Geologia:

1 QHa - Depósito Quaternário recente - Aluviões Fluviais Recentes
2 QHec - Cobertura Quaternário Holocênico Elúvio-Coluvial
3 TQm - Formação Murundu
4 TQt - Formação Taperinha
5 TQd - Cobertura Detrítica Colúvio-Eluvial (cascalho e canga)
6 Tdl - Cobertura detrito-laterítico (Terciário)
7 PMcr - Formação Cristo Redentor
8 PMg - Formação Ribeirão Gama

Classes da variável Geomorfologia:

1 Blocos_residuais
2 ModeladoDissecaçãoDiferencial
3 Modelado_ondulações_suaves
4 ModeladoPlano_Chapada
5 PlanoInclinadoColuvionado
6 Leques_de_Espraiamentos
7 Terraco_PlanicieAluvial
8 InundacoesPeriodicas_planiciesFluviais
9 InundacoesPeriodicas_exsudação_lencolfreatico
10 Depressão_material_fino

Classes da variável Solos:

1 CAMBISSOLO HÁPLICO Tb Distrófico
2 GLEISSOLO HÁPLICO Alumínico
3 GLEISSOLO HÁPLICO Tb Distrófico
4 GLEISSOLO MELÂNICO Alumínico
5 GLEISSOLO MELÂNICO Tb Distrófico
6 LATOSSOLO AMARELO Acriférrico

7 LATOSSOLO AMARELO Ácrico
8 LATOSSOLO VERMELHO Distrófico
9 LATOSSOLO VERMELHO Ácrico
10 LATOSSOLO VERMELHO-AMARELO Distrófico
11 LATOSSOLO VERMELHO-AMARELO Ácrico
12 NEOSSOLO LITÓLICO Distrófico
13 ORGANOSSOLO HÁPLICO Fíbrico
14 ORGANOSSOLO HÁPLICO Sáprico
15 PLINTOSSOLO HÁPLICO Distrófico
16 PLINTOSSOLO PÉTRICO Concrecionário
17 PLINTOSSOLO PÉTRICO Litoplíntico

Classes da variável-alvo Vegetação (classes-alvo):

1 Brejo
2 Brejo com Murundum
3 Campo Cerrado
4 Campo Limpo
5 Campo Limpo com Murundum
6 Campo Sujo
7 Campo Sujo com Murundum
8 Campo Úmido
9 Campo Úmido com Murundum
10 Cerrado
11 Cerradão
12 Floresta Galeria Paludosa
13 Floresta Galeria Seca
14 Vereda

As **classes de declividade**, geradas pela autora com as curvas de nível, foram "fatiadas" para as classes de declividade por percentagem de acordo com a Classificação da EMBRAPA (1979).

- Plano a suavemente ondulado (0 - 3%)
- Ondulado (3 - 8%)
- Forte ondulado (8 - 20%)
- Montanhoso (20 - 45%)
- Escarpado (> 45%)

5.4.2 Variáveis do Bioma Cerrado

Foram selecionadas 17 variáveis para o Bioma Cerrado, além da variável-alvo, totalizando 18 variáveis. Estas 17 variáveis foram agrupadas em 5 temas para facilitar as análises, em especial para a geração dos grafos das Redes Bayesianas. Muitas destas variáveis apresentavam alta cardinalidade (muitas classes), por isso não serão descritas aqui suas classes.

As 17 **variáveis preditoras** para o Bioma Cerrado foram:

- ZONA: zona climática
- TP_UMIDADE: tipo de umidade
- DISTR_UMID: distribuição de umidade no ano (meses sem chuva)
- TEMPERATUR: temperatura média
- nm_unid_gl: nome da unidade geológica
- nm_tempo_g: tempo geológico
- nm_dominio: nome do domínio geomorfológico
- nm_unidade: nome da unidade geomorfológica
- natureza: natureza do relevo
- caract: característica do relevo
- forma: forma do relevo
- compartime: compartimentação do relevo.
- solo_subgr: tipos de solo e subgrupos
- textura: textura do solo
- horizonte: horizonte A do solo
- AWC: água disponível no solo
- ClasseDecl: classes de declividade

Variável-alvo: nm_uveg - tipos de Savanas

Os agrupamentos de variáveis preditoras do Bioma Cerrado foram:

"Grupo_1" - **variáveis de clima**: 'ZONA', 'TP_UMIDADE', 'DISTR_UMID', 'TEMPERATUR';

"Grupo_2": **variáveis de relevo**: 'nm_dominio', 'nm_unidade', 'natureza', 'caract', 'forma', 'compartime';

"Grupo_3" – **variáveis de solo**: 'solo_subgr', 'textura', 'horizonte';

"Grupo_4" – **variáveis de geologia**: 'nm_unid_gl', 'nm_tempo_g';

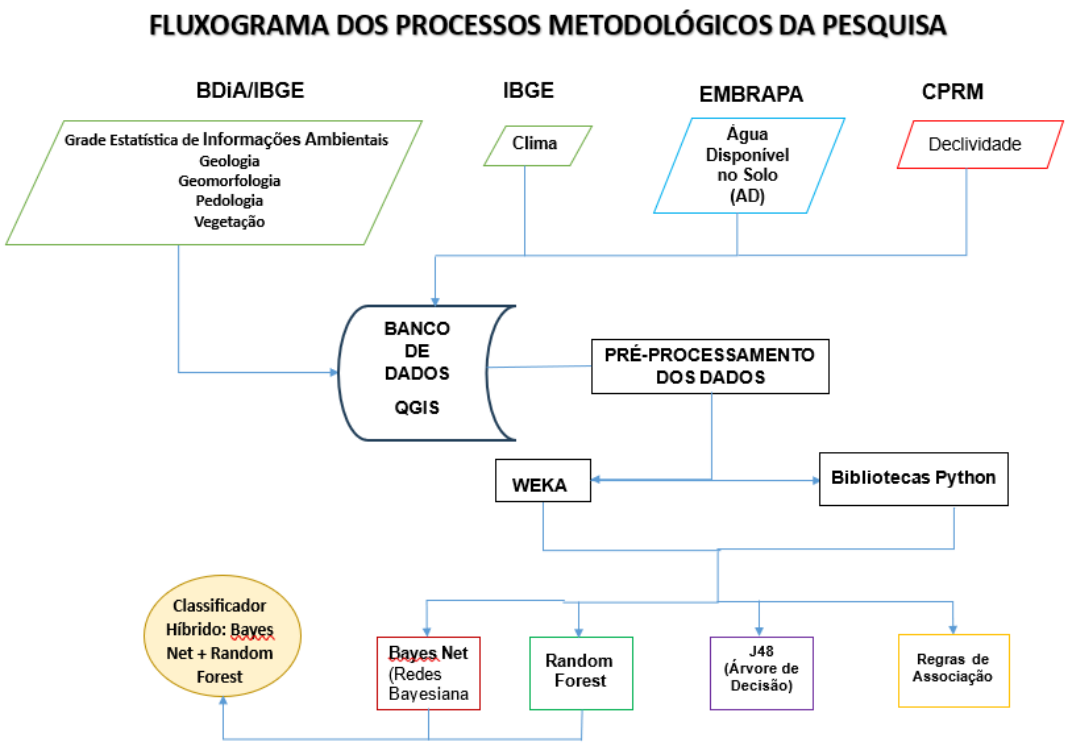
"Grupo_5" **variáveis relacionadas à água no solo**: 'ClasseAWC', 'ClasseDecl'.

5.5 Síntese dos processamentos

Todo o experimento foi implementado utilizando-se de dois dispositivos: um computador desktop institucional (Pós-Graduação do INPE) e um laptop de uso pessoal. O banco de dados foi estruturado no aplicativo QGIS versão 3.36 em um computador Intel(R) Core (TM) i7-2600 CPU @ 3.40GHz 3.40 GHz, 16 GB de RAM fabricado pela DELL Group, Sistema Operacional de 64 bits. Também foi utilizado como apoio em vários processamentos um laptop (computador portátil) DELL Intel(R) Core (TM) i7-8565U CPU @ 1.80GHz 1.9 GHz, 16 GB de RAM fabricado pela DELL Group, com Windows 11 Edição Home Single, Versão 24H2, Sistema Operacional de 64 bits e QGIS 10.2 instalados.

A Figura 5.3 mostra de forma sucinta todas as etapas da metodologia *ad hoc* realizadas no experimento.

Figura 5.3 – Esquema dos processamentos realizados na pesquisa.



Fonte: Elaborado pela autora.

Para a implementação de códigos Python em parte dos processamentos dos dados, nesta pesquisa, foram utilizados os modelos de linguagem de Inteligência Artificial (IA) DeepSeek-R1-Lite-Preview, da empresa chinesa DeepSeek; Inteligência Artificial Copilot, desenvolvida pela Microsoft; ChatGPT, desenvolvida pela OpenAI, e Sabiá 3, desenvolvida pela Maritaca AI (Brasil). Salienta-se que os modelos foram empregados para auxiliar na geração, otimização e análise de códigos, seguindo orientações específicas fornecidas pela autora para agilizar os testes de processamentos que não podiam ser feitos pelo aplicativo WEKA. A autora assume toda responsabilidade laboral dos códigos Python realizados pelos modelos IA's, uma vez que a mesma mantinha rigoroso controle nas demandas de cada etapa operacional da pesquisa a fim de garantir a reprodutibilidade da metodologia. Os resultados das IA's não podem substituir a avaliação crítica e a interpretação dos resultados feitos pela autora.

6 RESULTADOS E DISCUSSÃO

Assim como os processamentos, os resultados serão discutidos em duas etapas: os resultados da RECOR, na seção 6.1 e os resultados do Bioma Cerrado, na seção 6.2. No final do capítulo, será feita uma discussão síntese de tudo que foi alcançado para os objetivos propostos pela pesquisa.

6.1 Resultados da RECOR

Os processamentos realizados com os dados da RECOR foram: as regras de associação para as classes de vegetação do cerrado; o classificador J48 e o Random Forest. As classes de cerrado de interesse foram: **Campo Limpo, Campos Sujo, Campo Cerrado, Cerrado, Cerradão, Vereda e Floresta de Galeria**. Exemplos de relatórios das regras de associação geradas e de saídas gráficas das regras de associação feitas pelo visualizador de regras do WEKA estão no Apêndice C deste documento.

6.1.1 Regras de associação com os dados da RECOR

Inicialmente, as regras de associação foram geradas sem restrições, porém, observou-se que os subtipos de cerrado raramente apareciam como consequentes. Para compreender melhor as relações entre as variáveis ambientais e cada subtipo, optou-se por fixar a classe desejada como consequente (com a ferramenta CAR do WEKA), permitindo identificar associações relevantes que poderiam estar ocultas devido à baixa frequência dessas classes nas regras geradas no conjunto de dados.

Os valores de **Conviction** e **Confiança** foram filtrados para regras acima de 1,5 e 90%, respectivamente. Só foram consideradas as regras com *p-value* menor ou igual a 5% de significância estatística. Desta forma, as regras geradas foram forçadas a um resultado específico (classe de cerrado) de interesse, com controle estatístico robusto.

Na interpretação salienta-se que essas regras não são as mais fortes no conjunto de dados como um todo, mas sim dentro de uma análise dirigida.

6.1.1.1 Análise dos resultados da RECOR para as regras de associação, por cada classe de vegetação

a) CERRADÃO

Geologia:

- Tdl (Cobertura detrito-laterítico - Terciário): Presente em 100% das instâncias (807/807).

Geomorfologia:

- Modelado Plano Chapada: Presente em 100% das instâncias (807/807).

Solo:

LATOSSOLO VERMELHO Ácrico: Presente em 99,88% das instâncias (806/807).

Declividade:

- Plano (0 a 3%): Presente em 71,62% das instâncias (578/807).

- Suave Ondulado (3 a 8%): Presente em 28,38% das instâncias (229/807).

Resumo da Classe CERRADÃO:

A geologia Tdl (Cobertura detrito-laterítico - Terciário) e a geomorfologia Modelado Plano Chapada estão presentes em 100% das instâncias, sendo os atributos mais fortemente associados ao Cerradão. O solo LATOSSOLO VERMELHO Ácrico também é altamente relevante, presente em 99,88% das instâncias. A declividade Plano (0 a 3%) é a mais comum, presente em 71,62% das instâncias.

b) CERRADO

Geologia:

- Tdl (Cobertura detrito-laterítico - Terciário): Presente em 93,09% das instâncias (40.259/43.244).

Geomorfologia:

- Modelado Plano Chapada: Presente em 99,50% das instâncias (43.028/43.244).

Solo:

- LATOSSOLO VERMELHO Ácrico: Presente em 76,10% das instâncias (32.907/43.244).

Declividade:

- Plano (0 a 3%): Presente em 55,23% das instâncias (23.886/43.244).

- Suave Ondulado (3 a 8%): Presente em 43,35% das instâncias (18.745/43.244).

Resumo da Classe CERRADO:

A geomorfologia Modelado Plano Chapada é o atributo mais associado ao Cerrado, presente em 99,50% das instâncias. A geologia Tdl (Cobertura detrito-laterítico - Terciário) também é altamente relevante, com 93,09% de ocorrência. O solo LATOSSOLO VERMELHO Ácrico é o tipo de solo mais frequente, presente em 76,10% das instâncias. A maior parte das instâncias associadas ao "Cerrado" ocorre em áreas com declividade "Plano (0 a 3%)", representando 55,23% do total. Uma proporção significativa ocorre em áreas com declividade "Suave Ondulado (3 a 8%)", somando 43,35% das instâncias.

c) CAMPO CERRADO

Geologia:

- Tdl (Cobertura detrito-laterítico - Terciário): Presente em 56,40% das instâncias (10.980/19.469).

- TQd (Cobertura Detrítica Colúvio-Eluvial - cascalho e canga): Presente em 42,47% das instâncias (8.269/19.469).

Geomorfologia:

Modelado Plano Chapada: Presente em 89,43% das instâncias (17.410/19.469).

Solo:

- LATOSSOLO VERMELHO-AMARELO Ácrico: Presente em 34,43% das instâncias (6.704/19.469).

- LATOSSOLO AMARELO Ácrico: Presente em 21,50% das instâncias (4.185/19.469).

- PLINTOSSOLO PÉTRICO Concrecionário: Presente em 20,56% das instâncias (4.002/19.469).

Declividade:

- Suave Ondulado (3 a 8%): Presente em 54,88% das instâncias (10.685/19.469).

- Plano (0 a 3%): Presente em 37,31% das instâncias (7.264/19.469).

Resumo da Classe CAMPO CERRADO:

A geomorfologia Modelado Plano Chapada é o atributo mais associado ao Campo Cerrado, presente em 89,43% das instâncias. A geologia Tdl (Cobertura detrito-laterítico - Terciário) também é altamente relevante, com 56,40% de ocorrência. O solo LATOSSOLO VERMELHO-AMARELO Ácrico é o tipo de solo mais frequente, presente em 34,43% das instâncias. A declividade Suave Ondulado (3 a 8%) é a mais comum, presente em 54,88% das instâncias.

d) CAMPO SUJO

-Tdl (Cobertura detrito-laterítico - Terciário): Presente em 51,38% das instâncias (5.397/10.503).

-TQd (Cobertura Detrítica Colúvio-Eluvial - cascalho e canga): Presente em 41,45% das instâncias (4.353/10.503).

Geomorfologia:

-Modelado Plano Chapada: Presente em 70,05% das instâncias (7.357/10.503).

-Planos Inclinados Coluvionados: Presente em 20,45% das instâncias (2.148/10.503).

Solo:

-LATOSSOLO VERMELHO-AMARELO Ácrico: Presente em 28,73% das instâncias (3.018/10.503).

-LATOSSOLO AMARELO Ácrico: Presente em 23,39% das instâncias (2.456/10.503).

Declividade:

-Suave Ondulado (3 a 8%): Presente em 50,42% das instâncias (5.295/10.503).

-Plano (0 a 3%): Presente em 40,16% das instâncias (4.218/10.503).

Resumo da Classe CAMPO SUJO:

A geologia Tdl e a geomorfologia Modelado Plano Chapada são os atributos mais associados ao **Campo Sujo**, com 51,38% e 70,05% de ocorrência, respectivamente. O solo LATOSSOLO VERMELHO-AMARELO Ácrico é o tipo de solo mais frequente, presente em 28,73% das instâncias. A declividade Suave Ondulado (3 a 8%) é ligeiramente mais comum que a Plano (0 a 3%), com 50,42% e 40,16% de ocorrência, respectivamente.

e) VEREDA

Geologia:

-QHa (Depósito Quaternário recente - Aluviões Fluviais Recentes): Presente em 80,23% das instâncias (1.096/1.366).

Geomorfologia:

-Terraços Planícies Aluviais: Presente em 78,62% das instâncias (1.074/1.366).

Solo:

-ORGANOSSOLO HÁPLICO Sáprico: Presente em 81,92% das instâncias (1.119/1.366).

Declividade:

-Plano (0 a 3%): Presente em 50,73% das instâncias (693/1.366).

-Suave Ondulado (3 a 8%): Presente em 47,14% das instâncias (644/1.366).

Resumo da Classe VEREDA:

O solo ORGANOSSOLO HÁPLICO Sáprico é o atributo mais associado à **Vereda**, presente em 81,92% das instâncias. A geologia QHa (Aluviões Fluviais Recentes) e a geomorfologia Terraços Planícies Aluviais também são altamente relevantes, com 80,23% e 78,62% de ocorrência, respectivamente. A declividade "Plano (0 a 3%)" é ligeiramente mais comum que a "Suave Ondulado (3 a 8%)", com 50,73% e 47,14% de ocorrência, respectivamente.

6.1.1.2 Tabela síntese das regras de associação para a RECOR

A Tabela 6.1 apresenta uma síntese dos resultados: as classes de cerrado e os principais padrões associativos dos atributos, da RECOR, alcançados pelo algoritmo Apriori do WEKA durante o experimento.

Tabela 6.1 – Síntese das Regras de Associação para a RECOR.

Classe	Principais Associações dos Atributos
Vereda	Geologia: QHa (80,23%), Geomorfologia: Terraços Planícies Aluviais (78,62%), Solos: ORGANOSSOLO HÁPLICO Sáprico (81,92%), Declividade: Plano (0 a 3%) (50,73%)
Campo Cerrado	Geologia: Tdl (56,42%), TQd (42,46%), Geomorfologia: Modelado Plano Chapada (89,41%), Solos: LATOSSOLO VERMELHO-AMARELO Ácrico (34,43%), LATOSSOLO AMARELO Ácrico (21,49%), PLINTOSSOLO PÉTRICO Concrecionário (20,56%), Declividade: Suave Ondulado (3 a 8%) (54,89%), Plano (0 a 3%) (37,30%)
Cerrado	Geologia: Tdl (93,09%), Geomorfologia: Modelado Plano Chapada (99,50%), Solos: LATOSSOLO VERMELHO Ácrico (76,10%), Declividade: Plano (0 a 3%) (55,23%), Suave Ondulado (3 a 8%) (43,35%)
Campo Limpo	Geologia: TQd (50,00%), PMcr (29,80%), Geomorfologia: Planos Inclinados Coluvionados (41,39%), Modelado Dissecação Diferencial (24,59%), Modelado Plano Chapada (19,94%), Solos: CAMBISSOLO HÁPLICO Tb Distrófico (40,55%), LATOSSOLO AMARELO Ácrico (15,06%), LATOSSOLO VERMELHO-AMARELO Ácrico (10,71%), PLINTOSSOLO PÉTRICO Concrecionário (10,23%), Declividade: Suave Ondulado (3 a 8%) (59,89%), Ondulado (8 a 20%) (25,78%), Plano (0 a 3%) (13,37%)
Campo Sujo	Geologia: Tdl (51,38%), TQd (41,45%), Geomorfologia: Modelado Plano Chapada (70,05%), Planos Inclinados Coluvionados (20,45%), Solos: LATOSSOLO VERMELHO-AMARELO Ácrico (28,73%), LATOSSOLO AMARELO Ácrico (23,39%), Declividade: Suave Ondulado (3 a 8%) (50,42%), Plano (0 a 3%) (40,16%)
Cerradão	Geologia: Tdl (100%), Geomorfologia: Modelado Plano Chapada (100%), Solos: LATOSSOLO VERMELHO Ácrico (99,88%), Declividade: Plano (0 a 3%) (71,62%), Suave Ondulado (3 a 8%) (28,38%)

Fonte: Elaborado pela autora.

6.1.2 J48 para os dados da RECOR

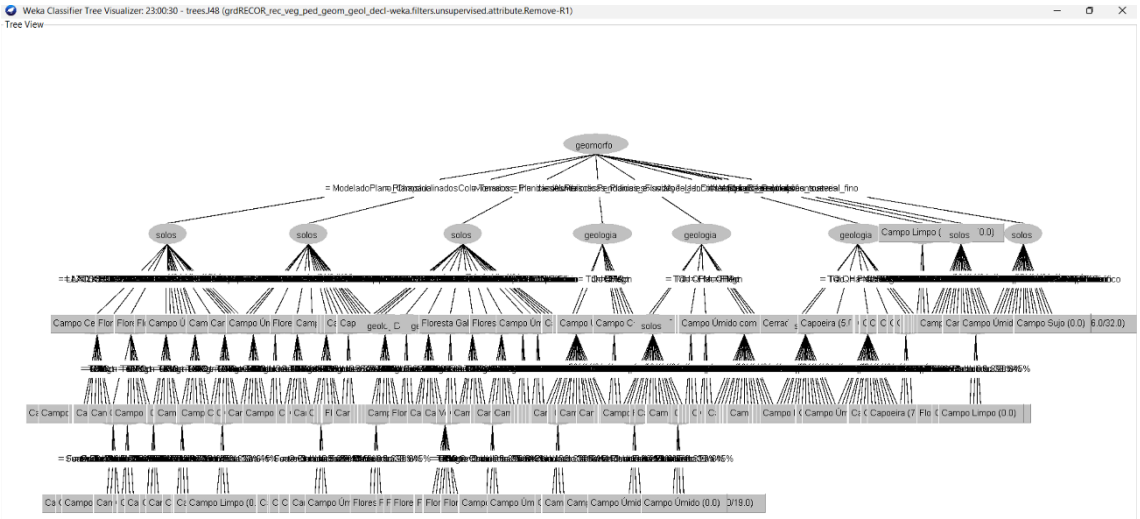
Os resultados do algoritmo J48 foram gerados no WEKA a partir de vários testes de configuração para achar o melhor tamanho da árvore de decisão. Depois de muitos testes, todos os resultados convergiram para uma média de 60% de acurácia global. Apesar de ser um resultado modesto para um classificador, é preciso levar em consideração que o conjunto de dados apresentavam apenas 4 variáveis preditoras e as classes de cerrado (variável alvo), para a RECOR, eram 15, ou seja, quase quatro vezes o número de variáveis para predição. Além disso, as classes eram extremamente desbalanceadas. Apesar do teste de SMOTE (*Synthetic Minority Over-sampling Technique*) para as classes minoritárias, o resultado da acurácia pouco se alterou. O SMOTE é uma técnica usada para lidar com conjuntos de dados desequilibrados, especialmente em aprendizado de máquina. Ele funciona criando amostras sintéticas da classe minoritária, em vez de simplesmente duplicar os exemplos existentes. O algoritmo utiliza os vizinhos mais próximos para gerar novos exemplos, combinando características dos dados originais. Entretanto, é preciso tomar muito cuidado para não embutir viés na criação de amostras sintéticas.

A classe cerradão foi extremamente problemática, pois nenhum algoritmo conseguiu discriminá-la com robustez estatística. Este resultado provavelmente se deveu à forma como o mapeamento de campo foi repassado para o banco de dados, pois a publicação do levantamento realizado na RECOR (IBGE, 2012, p. 155) discorre que houve discordância entre os autores sobre como classificar a área do cerradão. Para alguns, tratava-se de um cerrado denso, para outros especialistas temáticos tratava-se de Cerradão. A publicação descreve que houve um fenômeno observado em uma área descontínua de Cerrado Típico, no sul da RECOR, que exemplifica a transição do Cerrado Típico para o Cerradão. Na década de 1970, jovens indivíduos de espécies florestais (como *Copaifera langsdorffii* e *Siphoneugenia densiflora*) surgiram ao longo de uma antiga estrada de tropeiros. Essas árvores cresceram, reproduziram-se e aumentaram significativamente a densidade e diversidade da vegetação local. A transformação parece ter sido favorecida pelo solo profundo e argiloso, ausência de incêndios há mais de três décadas e presença de animais dispersores de

sementes. Este fato pode explicar porque todos os algoritmos apresentaram confusão com a classe “cerrado” pois os autores optaram por considerar, no levantamento, como cerrado (típico).

Para se ter uma ideia de como o algoritmo organizou as regras para a classificação das variáveis-alvo, com os atributos fornecidos, a Figura 6.1 mostra a saída da Árvore de Decisão, feita pelo WEKA, para os dados da RECOR.

Figura 6.1 – Visão da Árvore de Decisão (algoritmo J48) para os dados da RECOR.



Fonte: Elaborado pela autora.

Interessante notar que para este algoritmo, o nó principal de onde saem os ramos tem como critério a variável relacionada à forma do relevo. Neste conjunto de dados não havia dados de clima. Esta configuração, onde o relevo se destacava entre as variáveis preditoras também foi observado em outros algoritmos testados.

6.1.3 RF para os dados da RECOR

A execução de testes de classificação com o RF, para a RECOR, demonstrou que o número de árvores da floresta não precisava ser alto para ter um bom desempenho. O número de instâncias permitiu a execução com 100 árvores, mas o melhor valor de acurácia foi obtido pela execução do algoritmo com apenas 15 árvores. Também foram testados outros algoritmos para a RECOR, no WEKA além dos mencionados no presente documento, entretanto, como foi

dito, a acurácia de todos os algoritmos testados resultou em torno de 60%. Não mais que isso.

6.2 Resultados do Bioma Cerrado

Os resultados obtidos para o Bioma Cerrado foram feitos para as Regras de Associação, algoritmo J48, algoritmo RF, algoritmo BN e para o modelo híbrido RF + BN. As regras de associação foram feitas com o algoritmo Apriori do WEKA. Também foram gerados no WEKA os relatórios de desempenho, porém, algumas etapas foram feitas com ajuda das bibliotecas Python.

6.2.1 A análise MCA para as savanas do Bioma Cerrado

A fim de entender a variabilidade das características das Savanas foi feito um gráfico de Análise de Correspondência Múltipla (MCA - *Multiple Correspondence Analysis*). A análise MCA é uma técnica estatística usada para explorar e visualizar a estrutura de associações entre várias variáveis categóricas em um conjunto de dados.

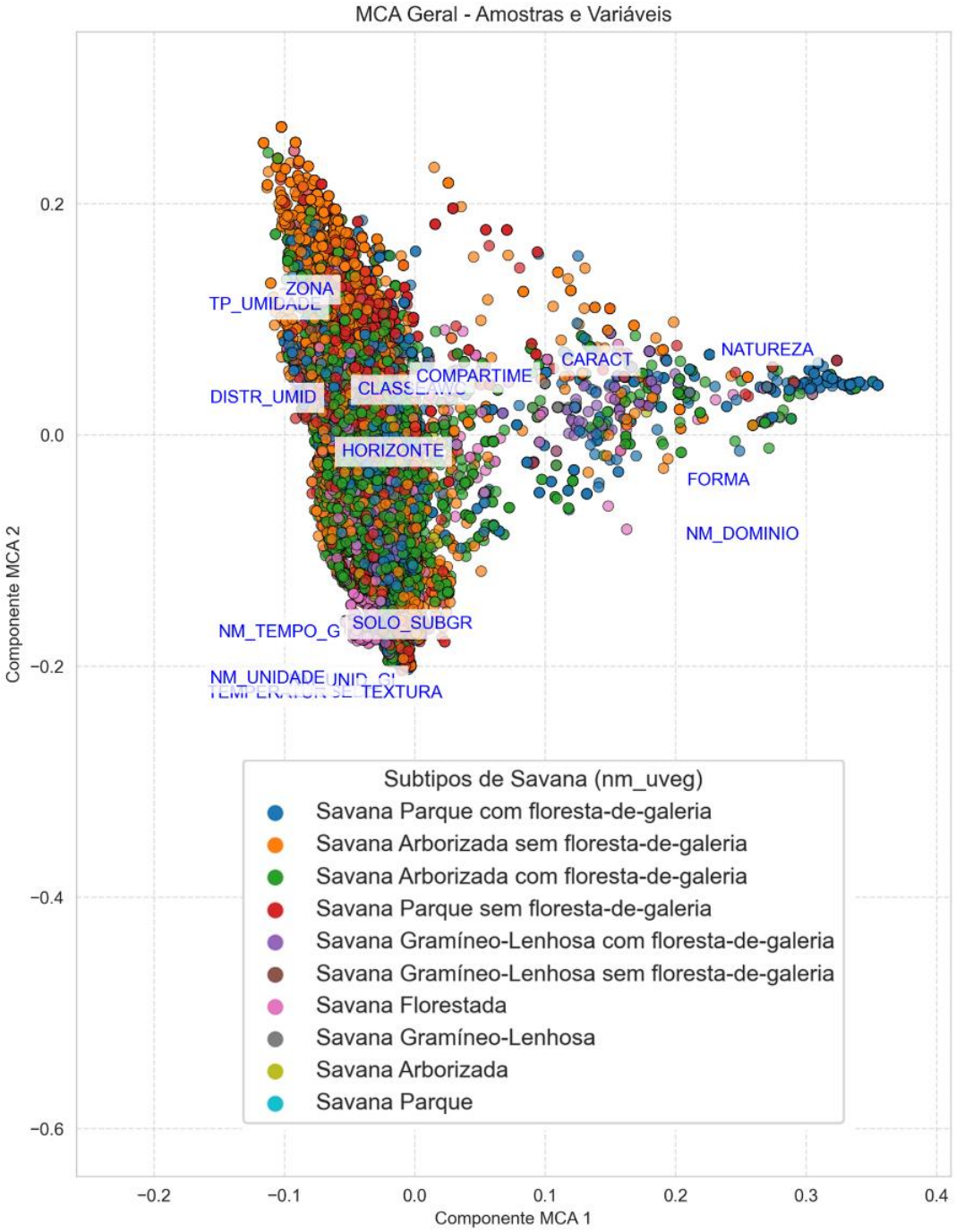
O MCA transforma variáveis categóricas em um espaço de dimensões reduzidas, permitindo representar essas variáveis em um gráfico visual, facilitando a interpretação dos padrões e associações entre elas. A técnica busca identificar dimensões principais que expliquem a variação nos dados, com base na proximidade entre categorias ou níveis das variáveis no gráfico resultante.

Neste experimento, devido à natureza e volume dos dados, foi realizado através de Biblioteca Python, um gráfico das variáveis preditoras para uma amostragem estratificada de 10%, a fim de entender como estavam se comportando as variáveis preditoras em relação à variável alvo, na amostra (Figura 6.2).

O gráfico demonstrou aquilo que visualmente os especialistas de campo percebem: como as savanas apresentam características em comum. Isto é perceptível no gráfico MCA pela alta concentração de amostras e variáveis próximo à origem (0,0). Na componente principal (eixo X) consegue-se perceber o distanciamento das variáveis relacionadas ao grupo de relevo, demonstrando que esses atributos fornecem elementos para a diversificação das Savanas.

Em relação ao eixo y, segunda componente principal, destaca-se que as variáveis relacionadas ao clima formam um agrupamento na parte superior esquerda, evidenciando que algumas características das savanas também podem ser diferenciadas pelo clima.

Figura 6.2 – Análise MCA para os tipos de Savana do Bioma Cerrado.



Fonte: Elaborado pela autora.

Esta análise dá uma dimensão de como também é difícil para algoritmos classificadores discriminarem os subtipos de savanas. Assim, a pesquisa foi direcionada para saber qual(is) seria(m) o(s) algoritmo(s) com melhor desempenho para classificar um conjunto de amostras com características muito próximas (semelhantes) com relação às variáveis preditoras (atributos das savanas). Esta questão também era um dos objetivos do trabalho.

6.2.2 Regras de associação para os dados do Bioma Cerrado

As regras de associação para o Bioma Cerrado, no WEKA, assim como na RECOR não apresentaram espontaneamente alguma relação com as classes de interesse. Também foi usada a ferramenta CAR do WEKA para forçar as regras de interesse, ou seja, com o consequente para as classes de savana.

6.2.2.1 **Tabela síntese das regras de associação para o Bioma Cerrado**

Os resultados das Regras de Associação para o Bioma Cerrado, estão expressos na Tabela 6.2, com as regras dirigidas[4] com os mesmos critérios feitos para a RECOR, ou seja, com os mesmos critérios de métricas estatísticas mencionados na Seção 6.1.

[4] A grafia de muitos termos que constam no BDiA aparece ainda da forma como eram escritos antes do Acordo Ortográfico da Língua Portuguesa de 1990 — que começou a ser implementado oficialmente em países lusófonos a partir de 2009. A palavra “semiúmida” (ou “semiúmido”), por exemplo, aparece escrita como “semi-úmida” (ou “semi-úmido”), pois era comum ver essas palavras com uso de hífen, uma vez que os dados foram preenchidos anteriormente à Reforma Ortográfica. Na elaboração do código em Python, foi utilizado a grafia que consta no Banco de Dados, a fim de fazer a consulta e realizar a busca das informações necessárias para as regras de associação.

Tabela 6.2 – Síntese dos principais padrões associativos dos atributos pelas Regras de Associação.

Subtipo de Savana	Principais Associações dos Atributos
Savana Florestada	• Zona: Tropical Brasil Central • Horizonte: A moderado • Temperatura: quente (>18°C todos os meses) • Água Disponível no solo: AD5 • Natureza do relevo: dissecação • Característica do relevo: dissecação homogênea/diferencial • Domínio: Bacias e Coberturas Sedimentares Fanerozoicas
Savana Parque	• Horizonte: A moderado • Compartimento: Planaltos (Planalto dos Parecis) • Domínio: Bacias e Coberturas Sedimentares Fanerozoicas • Temperatura: quente (>18°C) • Distribuição umidade: 3 meses secos • Tipo umidade: úmido • Zona: Equatorial
Savana Parque com Floresta-de-Galeria	• Temperatura: quente (>18°C) • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Água Disponível no solo: AD5 • Zona: Tropical Brasil Central • Horizonte: A moderado
Savana Parque sem Floresta-de-Galeria	• Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Temperatura: quente (>18°C) • Zona: Tropical Brasil Central • Água Disponível no solo: AD5
Savana Arborizada	• Zona: Tropical Brasil Central • Água Disponível no solo: AD5 • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Horizonte: A moderado
Savana Arborizada sem Floresta-de-Galeria	• Zona: Tropical Brasil Central • Temperatura: quente (>18°C) • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Domínio: Bacias e Coberturas Sedimentares Fanerozoicas

continua

Tabela 6.2 – Conclusão.

Subtipo de Savana	Principais Associações dos Atributos
Savana Arborizada com Floresta-de-Galeria	• Temperatura: quente (>18°C) • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Água Disponível no solo: AD5 • Zona: Tropical Brasil Central • Horizonte: A moderado • Natureza do relevo: dissecação • Característica do relevo: dissecação homogênea/diferencial
Savana Gramíneo-Lenhosa	• Zona: Tropical Brasil Central • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Água Disponível no solo: AD5 • Temperatura: quente (>18°C) • Horizonte: A moderado • Domínio: Crátons Neoproterozoicos
Savana Gramíneo-Lenhosa com Floresta-de-Galeria	• Água Disponível no solo: AD5 • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Temperatura: quente (>18°C) • Natureza do relevo: dissecação • Característica do relevo: dissecação homogênea/diferencial • Horizonte: A moderado
Savana Gramíneo-Lenhosa sem Floresta-de-Galeria	• Água Disponível no solo: AD5 • Zona: Tropical Brasil Central • Tipo umidade: semi-úmido • Distribuição umidade: 4-5 meses secos • Domínio: Cinturões Móveis Neoproterozoicos • Natureza do relevo: dissecação • Temperatura: subquente (15-18°C em ≥1 mês)

Fonte: Elaborado pela autora.

Os resultados das regras de associação, para o Bioma Cerrado, destacaram a recorrência de variáveis como horizonte A moderado, Zona Tropical Brasil Central e Classe AD5 de água no solo, que aparecem como fatores comuns em múltiplos subtipos de savana, evidenciando seu papel fundamental na caracterização dessas formações vegetacionais. As análises revelaram distinções marcantes entre os subtipos, particularmente no caso da Savana Parque, que se diferencia por apresentar apenas 3 meses secos em contraste com os 4-5 meses característicos dos demais subtipos, e da Savana Gramíneo-Lenhosa sem floresta-de-galeria, única com ocorrência em condições de

temperatura subquente (15-18°C em pelo menos um mês). Esses padrões demonstram como arranjos específicos de variáveis climáticas (temperatura e distribuição de umidade), pedológicas (disponibilidade hídrica no solo) e geomorfológicas atuam sinergicamente na diferenciação dos subtipos de savana. A presença ou ausência de floresta-de-galeria emergiu como um elemento discriminante relevante, associado principalmente a variações nos regimes hídricos e nas características do substrato geológico, reforçando a importância desses fatores ambientais na configuração das paisagens savânicas.

6.2.2.2 Padrões gerais nas associações de Savanas do Cerrado

As savanas analisadas compartilham horizonte A moderado, classe de água disponível no solo AD5 e localização no Brasil Central, mas divergem em umidade e geologia. Savanas com floresta-de-galeria exigem 4 a 5 meses secos e umidade semiúmida, enquanto as sem floresta vinculam-se a domínios geológicos específicos (bacias ou cinturões). A exceção é a Savana Gramíneo-Lenhosa sem floresta, única com temperatura subquente.

6.2.3 J48 para os dados do Bioma Cerrado

Os resultados de desempenho do classificador e por classes dos algoritmos J48 para o Bioma Cerrado estão contemplados nas Tabela 6.3 e Tabela 6.4.

Tabela 6.3 – Resultado das métricas de desempenho do classificador J48.

Métrica	Valor
Instâncias Corretamente Classificadas	379.708 (79,16%)
Instâncias Incorretamente Classificadas	99.940 (20,84%)
Kappa Statistic	0,7315
Erro Absoluto Médio	0,0602
Erro Quadrático Médio	0,174
Erro Absoluto Relativo	38,58%
Erro Quadrático Relativo	62,32%
Total de Instâncias	479.648

Fonte: Elaborado pela autora.

Os resultados apresentados pelo J48 foram considerados bons, entretanto quando analisado pelo desempenho por classes (Tabela 6.4), é notório a dificuldade de discriminar a Savana Parque, pelo fato desta classe ser minoritária. Ela é sub-representada quando se avalia o quantitativo dela em relação às demais.

Para minimizar o efeito do desbalanceamento foi aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), através do filtro do WEKA, porém os ganhos de desempenho na classificação foram mínimos.

Tabela 6.4 – Desempenho do algoritmo classificador J48 para as classes do Bioma Cerrado.

Classe	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Savana Parque com floresta-de-galeria	0.788	0.043	0.826	0.788	0.806	0.758	0.961	0.869
Savana Arborizada com floresta-de-galeria	0.841	0.099	0.782	0.841	0.811	0.727	0.944	0.881
Savana Parque sem floresta-de-galeria	0.701	0.041	0.752	0.701	0.726	0.679	0.949	0.806
Savana Florestada	0.624	0.010	0.704	0.624	0.662	0.651	0.977	0.735
Savana Arborizada sem floresta-de-galeria	0.813	0.066	0.807	0.813	0.810	0.746	0.957	0.884
Savana Gramíneo-Lenhosa com floresta-de-galeria	0.812	0.009	0.787	0.812	0.799	0.791	0.991	0.869
Savana Gramíneo-Lenhosa sem floresta-de-galeria	0.816	0.001	0.884	0.816	0.849	0.848	0.997	0.920
Savana Gramíneo-Lenhosa	0.590	0.000	0.838	0.590	0.692	0.702	0.996	0.781
Savana Arborizada	0.818	0.001	0.785	0.818	0.801	0.801	0.998	0.856
Savana Parque	0.000	0.000	?	0.000	?	?	0.995	0.187

Fonte: Elaborado pela autora.

Considerando que a Árvore de Decisão (J48) apresentou algumas dificuldades de desempenho para a classe minoritária Savana Parque, mesmo com a utilização do filtro SMOTE (notar que devido à ausência de acertos, aparece o

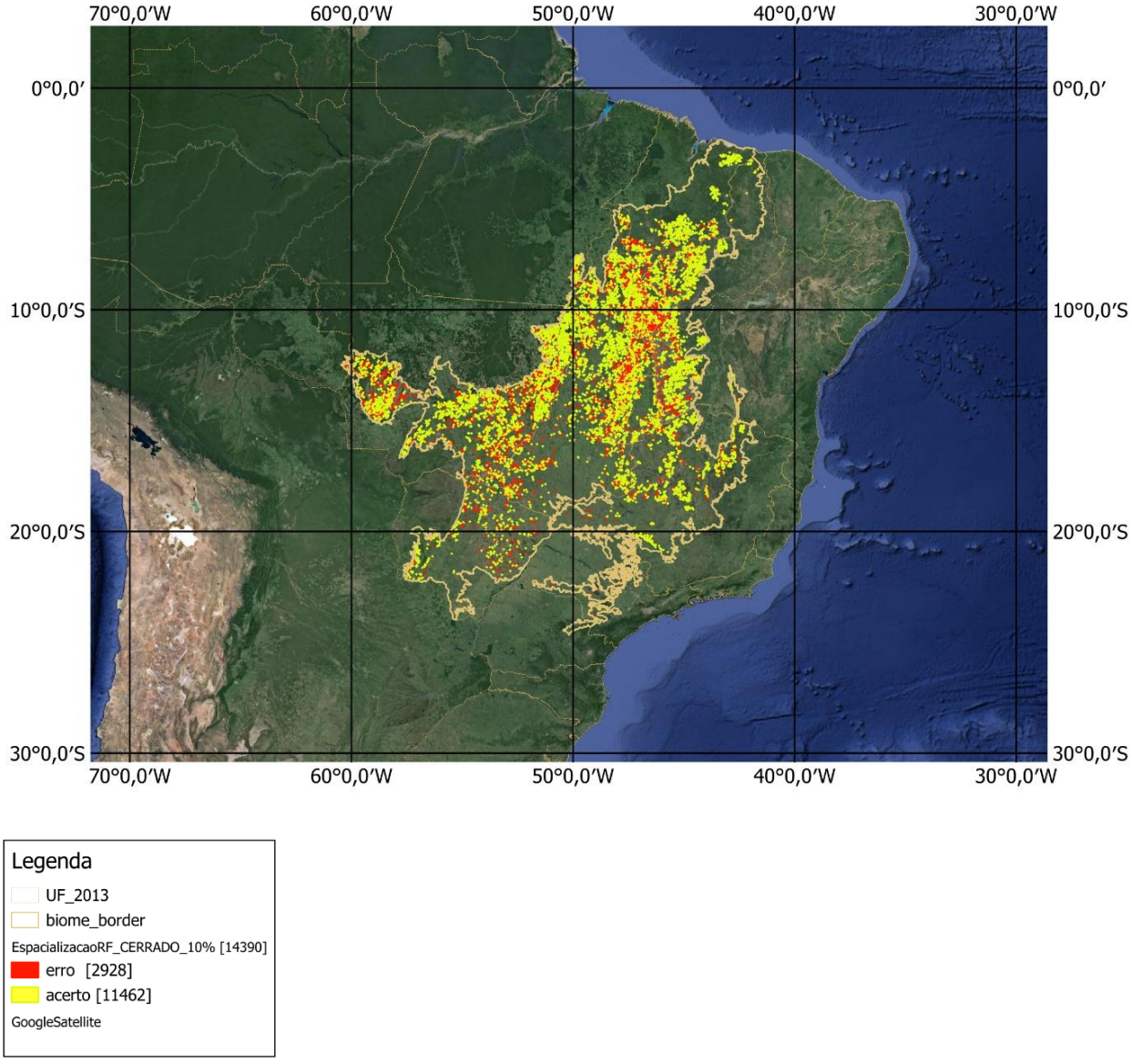
sinal de “?” na Tabela 6.4), o experimento passou então para a utilização do classificador RF.

6.2.4 RF para os dados do Bioma Cerrado

Os Classificadores Random Forests geralmente superam árvores de decisão individuais em termos de precisão e capacidade de generalização. O problema do algoritmo é que, justamente por se tratar de uma floresta de árvores de decisão, ele exige um desempenho de processamento bem mais pesado. No caso dos dados da RECOR, na casa de centena de milhar, o número de instâncias não foi impedimento para processar o classificador. Mas para os dados do Bioma Cerrado, foi necessário trabalhar com amostragens porque o aplicativo WEKA não conseguia processar todo o conjunto de dados. Então, optou-se por amostragens estratificadas.

A amostragem estratificada foi feita, em código Python e no WEKA. Com as amostras estratificadas geradas pelo código Python, foi possível também gerar a saída do classificador para espacialização no QGIS (Figura 6.3).

Figura 6.3 – Espacialização do algoritmo Random Forest para o Bioma Cerrado (amostragem de 10%).



Fonte: Elaborado pela autora.

Visualmente é possível observar o desempenho robusto do RF na discriminação das classes de Savanas. Numericamente, de um total de 14.390 instâncias com classes de Savanas, houve um acerto de 11.492 instâncias (células amarelas) corretamente classificadas pelo RF, ou seja 79,65% de acerto nesta amostragem. Entretanto, ao analisar os resultados de desempenho do algoritmo RF para cada classe de Savana, apresentados na Tabela 6.5, percebe-se que o

algoritmo não discrimina muito bem a classe minoritária Savana Parque, assim como o J48. Também foram testados os filtros de SMOTE nos dados, através do WEKA, mas o SMOTE contribuiu muito pouco para melhorar a acurácia da classificação.

Diante deste desafio, o experimento buscou alternativas para melhorar o desempenho de classes tão desbalanceadas, como as Savanas Parque, no conjunto de dados.

Tabela 6.5 – Desempenho por classe de Bioma Cerrado para o algoritmo RF (50% de amostragem estratificada).

Classe	TP Rate (Recall)	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Savana Parque com floresta-de-galeria	0,835	0,038	0,849	0,842	0,802	0,975	0,913
Savana Arborizada com floresta-de-galeria	0,866	0,071	0,837	0,851	0,787	0,966	0,929
Savana Parque sem floresta-de-galeria	0,755	0,033	0,802	0,778	0,740	0,968	0,874
Savana Florestada	0,741	0,007	0,787	0,764	0,756	0,983	0,849
Savana Arborizada sem floresta-de-galeria	0,861	0,054	0,843	0,852	0,801	0,973	0,928
Savana Gramíneo-Lenhosa com floresta-de-galeria	0,859	0,006	0,858	0,858	0,852	0,993	0,930
Savana Gramíneo-Lenhosa sem floresta-de-galeria	0,888	0,001	0,925	0,906	0,905	0,997	0,964
Savana Gramíneo-Lenhosa	0,796	0,000	0,918	0,853	0,855	0,994	0,925
Savana Arborizada	0,878	0,000	0,924	0,900	0,900	0,993	0,943
Savana Parque	0,282	0,000	0,310	0,295	0,295	0,996	0,361

Fonte: Elaborado pela autora.

6.2.5 BayesNet (BN) para os dados do Bioma Cerrado

As Redes Bayesianas modelam relações probabilísticas entre variáveis, levando em conta dependências condicionais. Isso poderia ajudar a capturar padrões específicos das variáveis que caracterizam a Savana Parque. Apesar da acurácia geral mais baixa, o algoritmo Bayesian Networks ou BayesNet, quando

comparado com os modelos anteriores J48 e RF, apresentou melhor desempenho para discriminar a "Savana Parque". Provavelmente porque o BN consegue lidar melhor com classes minoritárias ao calcular probabilidades diretamente (Tabela 6.6). Já o Random Forest tende a favorecer classes mais frequentes, o que pode explicar porque a Savana Parque foi mal classificada, quando comparada com as demais classes de Savana.

Tabela 6.6 - Desempenho por classe do classificador BN para amostragem de 50% do Bioma Cerrado.

Classe	**TP Rate (Recall)**	**FP Rate**	**Precision**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
Savana Arborizada com floresta-de-galeria	0,673	0,161	0,640	0,656	0,505	0,848	0,709
Savana Parque sem floresta-de-galeria	0,628	0,092	0,548	0,585	0,507	0,889	0,659
Savana Florestada	0,689	0,043	0,372	0,484	0,483	0,940	0,527
Savana Arborizada sem floresta-de-galeria	0,615	0,079	0,725	0,666	0,568	0,887	0,755
Savana Gramíneo-Lenhosa com floresta-de-galeria	0,760	0,030	0,516	0,615	0,607	0,964	0,695
Savana Gramíneo-Lenhosa sem floresta-de-galeria	0,817	0,007	0,521	0,636	0,648	0,984	0,801
Savana Gramíneo-Lenhosa	0,790	0,008	0,244	0,373	0,436	0,982	0,602
Savana Arborizada	0,818	0,003	0,554	0,661	0,672	0,996	0,779
Savana Parque	0,984	0,002	0,177	0,299	0,416	0,999	0,275

Fonte: Elaborado pela autora.

Uma grande vantagem do classificador BN é que ele gera as Tabelas de Distribuição de Probabilidade (TDP) ao criar os grafos (exemplos nos Apêndices D e E, no final do trabalho). As TDPs representam as probabilidades condicionais de uma variável, dado o estado de suas variáveis parentais (ou antecedentes) na rede. Cada nó na rede Bayesiana possui uma TDP associada, que descreve como a variável depende de seus pais. As TDPs permitem calcular a

probabilidade de eventos ou estados desconhecidos com base em evidências observadas.

No experimento, foram geradas as tabelas e confeccionados os respectivos gráficos para a as Savanas, a fim de saber quais atributos estão relacionados com sua ocorrência, para um valor acima de 50% de probabilidade. A Figura 6.4 traz esta informação que se traduz em uma síntese do que foi buscado durante o experimento, considerando uma amostragem estratificada de 10 %.

Para facilitar o entendimento, o gráfico foi colorido de acordo com o grupo de variáveis descrito na Seção 5.4.2. Desta forma, percebe-se o predomínio de variáveis relacionadas ao clima (cor vermelha) e as variáveis relacionadas ao relevo (cor azul) de atributos naturais na ocorrência das Savanas, como um todo, no Bioma Cerrado.

O atributo ambiental que aparece em primeiro lugar é o predomínio de regiões em Bacias e Coberturas Sedimentares Fanerozoicas, seguido de temperatura média quente (acima de 18°C em todos os meses do ano. Tipo de umidade úmido, a disponibilidade de água (AD5) entre 100-200 cm de profundidade, novamente o tipo de umidade (semiúmida), distribuição de umidade (4 a 5 meses seco), compartimentação em Planaltos, distribuição de chuvas (3 meses secos), entre outros atributos.

Estes resultados refletem como o clima e o relevo apresentam papel preponderante na resposta da vegetação frente às mudanças climáticas. Vários estudos com abordagem semelhantes investigaram como condições ambientais determinam a composição de múltiplos biótopos em uma mesma área. No que diz respeito aos resultados de clima, o trabalho de Rubanschi et al. (2023), ao comparar clima e solos, demonstrou que é o clima o principal limitante ambiental para a distribuição de biótopos, enquanto as condições de solo são limitantes em contextos específicos.

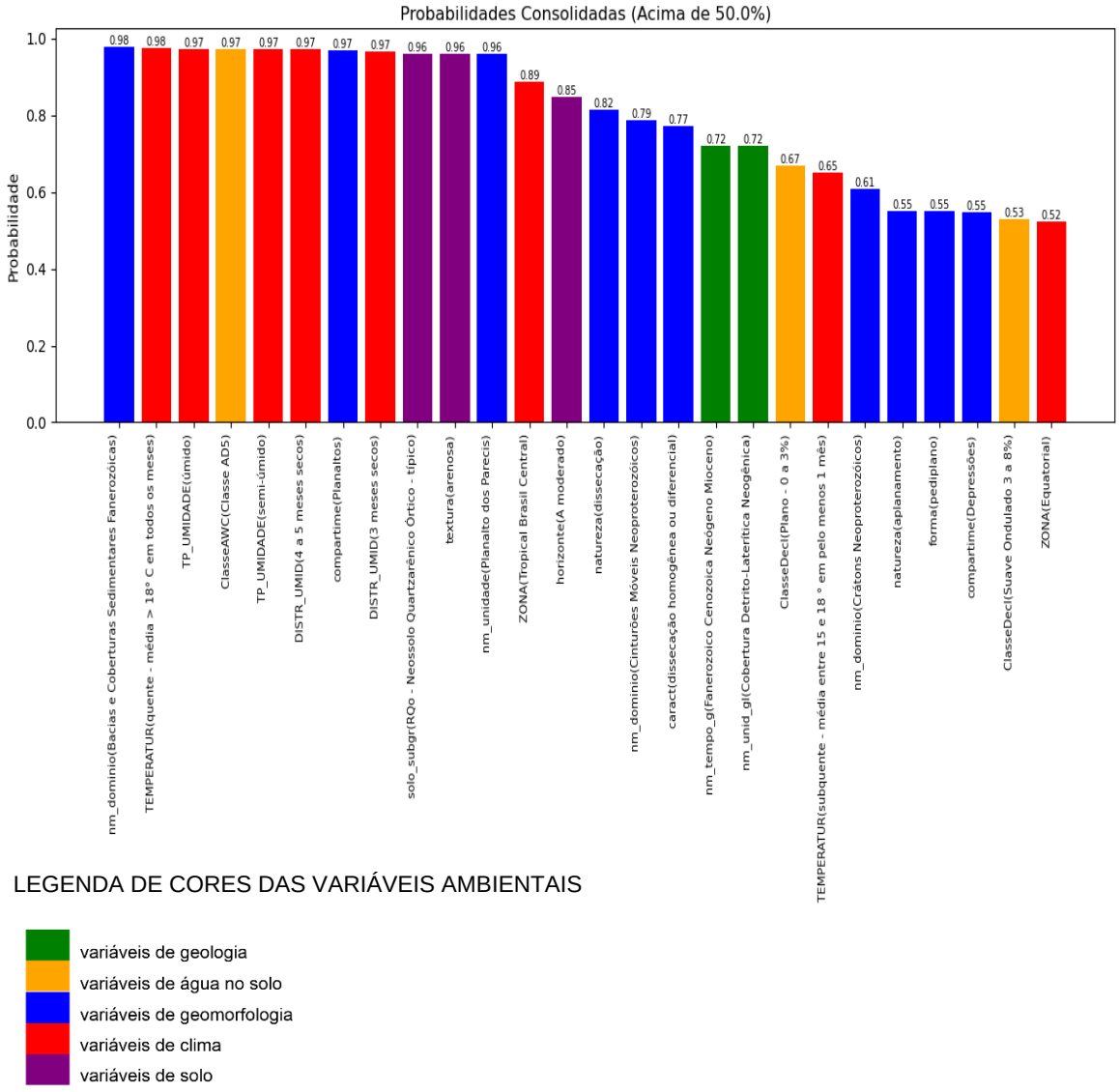
Com relação a classificação da vegetação e as formas de relevo, Kuplich et al. (2007) citam os trabalhos pioneiros de levantamento de recursos naturais, do projeto RADAM, na década de 70, para a classificação florística apoiados em imageamento de radar. Para Kuplich et al. (2007), as vantagens da predição da

vegetação com base na topografia ou variáveis geomorfométricas derivadas residem na possibilidade de fornecer dados alternativos, confiáveis e, sobretudo, dotados de significado ecofisiológico, para o estabelecimento ou refinamento do conhecimento acerca da vegetação, elementos essenciais na avaliação e determinação de prioridades de conservação.

Um aspecto complementar a ressaltar é que no grafo de rede bayesiana, considerando todas as variáveis ambientais utilizadas no estudo, a variável que centralizou todas as relações de **causalidade**, preponderando entre as demais foi a que expressa a unidade geomorfológica. Provavelmente pelo fato de que o nome da unidade geomorfológica, no Banco de Dados de Informações Ambientais do IBGE, considera várias toponímias e estas expressam, por sua vez, várias condições ambientais, tais como: nome de regiões hidrográficas, nomes de serras, entre outras informações, que por si só, agregam várias condicionantes naturais que influenciam a ocorrência de tipos vegetacionais.

Os resultados encontrados neste experimento devem ser submetidos à análise de especialistas temáticos de vegetação, para que possam avaliar até que ponto a metodologia é compatível com a escala do que foi levantado em campo. O ganho de conhecimento advindo com o aprendizado de máquina e mineração de dados deve sempre ser confrontado com o conhecimento especializado. Este é o fluxo metodológico do KDD (Fayyad et al., 1996).

Figura 6.4 – Probabilidades consolidadas, por variável ambiental para ocorrência de Savanas no Bioma Cerrado (valores acima de 50%).



Fonte: Elaborado pela autora.

6.2.6 Classificador híbrido (RF + BN) para o Bioma Cerrado

Em busca de alcançar os objetivos da pesquisa em conseguir um modelo classificador que discriminasse todas as savanas do Bioma Cerrado, a partir dos padrões associativos entre as variáveis ambientais do experimento, optou-se finalmente por um classificador híbrido. O classificador deveria capturar tão bem

as classes majoritárias, como o Random Forest e, também, as minoritárias, como o BayesNet.

Assim, através do aplicativo Weka, foi criado um classificador híbrido usando o meta-classificador “Vote”, que combina as previsões de dois algoritmos diferentes, no caso, o Random Forest (RF) e o BayesNet (BN). Assim, cada classificador RF e o BN faz suas previsões independentemente. O meta-classificador “Vote” combina essas previsões calculando a média das probabilidades de cada classe. A classe final é determinada pela maior probabilidade média. Essa abordagem híbrida aproveita as forças de ambos os algoritmos: a capacidade do RF de lidar com relações complexas nos dados e a abordagem probabilística do BayesNet.

Os desempenhos individuais dos algoritmos RF e BN estão representados na Tabela 6.7.

Tabela 6.7 – Comparação entre os desempenhos individuais dos classificadores Random Forest e Bayes Net para os dados do Bioma Cerrado.

Métrica	Random Forest (RF)	Bayesian Networks (BN)
Instâncias Corretamente Classificadas (%)	83,65%	63,41%
Instâncias Incorretamente Classificadas (%)	16,35%	36,59%
Kappa Statistic	0,7897	0,5385
Mean Absolute Error (MAE)	0,0452	0,0764
Root Mean Squared Error (RMSE)	0,1526	0,2454
Relative Absolute Error (RAE)	29,01%	48,97%
Root Relative Squared Error (RRSE)	54,66%	87,89%
Total de Instâncias	239884	239884

Fonte: Elaborado pela autora.

Analisando o quadro comparativo, a acurácia do Bayesian Networks (BN) foi 63,41%, enquanto a do Random Forest (RF) foi 83,65%, ou seja, o RF teve um desempenho significativamente melhor na classificação geral. O classificador BN

conseguiu discriminar melhor a classe "Savana Parque" (TP Rate = 0,984 vs. 0,282 no RF), provavelmente porque utiliza relações probabilísticas que podem favorecer classes menores.

O relatório de desempenho completo, por classe, do classificador híbrido RF + BN está apresentado na Tabela 6.8.

Tabela 6.8 - Precisão Detalhada Por Classe do Modelo Hibrido RF + BN.

Classe	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Savana Parque com floresta-de-galeria	0,629	0,029	0,849	0,629	0,723	0,675	0,965	0,870
Savana Arborizada com floresta-de-galeria	0,784	0,138	0,706	0,784	0,743	0,628	0,931	0,859
Savana Parque sem floresta-de-galeria	0,673	0,055	0,685	0,673	0,679	0,623	0,945	0,790
Savana Florestada	0,720	0,030	0,470	0,720	0,569	0,563	0,976	0,709
Savana Arborizada sem floresta-de-galeria	0,732	0,073	0,771	0,732	0,751	0,670	0,951	0,866
Savana Gramíneo-Lenhosa com floresta-de-galeria	0,823	0,017	0,672	0,823	0,740	0,732	0,989	0,855
Savana Gramíneo-Lenhosa sem floresta-de-galeria	0,820	0,004	0,666	0,820	0,735	0,736	0,997	0,897
Savana Gramíneo-Lenhosa	0,818	0,003	0,496	0,818	0,618	0,636	0,998	0,815
Savana Arborizada	0,837	0,001	0,691	0,837	0,757	0,760	0,999	0,909
Savana Parque	1,000	0,002	0,178	1,000	0,302	0,421	1,000	0,420
Weighted Avg.	0,723	0,076	0,738	0,723	0,724	0,651	0,950	0,847

Fonte: Elaborado pela autora.

A hibridização conseguiu melhorar significativamente o recall para a classe Savana Parque, passando de 0,282 no RF puro para 1,000 no modelo híbrido. Isso significa que o híbrido capturou todas as instâncias dessa classe, o que é uma grande vantagem quando se trabalha com classes minoritárias que não respondem bem ao SMOTE, como foi o caso de nosso conjunto de dados.

Mesmo com Precision baixa, o MCC (Matthews Correlation Coefficient) do híbrido (0,421) é superior ao RF (0,295) e J48 (?)[5], indicando um equilíbrio melhor entre Recall e Precision.

A pesquisa optou por priorizar Recall para Savana Parque, mesmo com Precision baixa, porque o custo ecológico de ignorar áreas reais de Savana Parque pode ser irreversível quando o objetivo é mapear áreas potenciais de cerrado (savanas) para fins de preservação ou restauração.

Diante desses resultados, o experimento foi finalizado com a escolha de um classificador híbrido, sendo parte com Random Forest e parte com BN, a fim de poder encontrar um desempenho que consiga discriminar classes minoritárias, bem como ter robustez para as classes majoritárias também.

É importante ressaltar que houve perdas de acurácia no modelo híbrido, mas priorizar o Recall, na pesquisa, foi justificável e estratégico, pois falsos negativos (não detectar Savana Parque) seriam críticos para a preservação desta tipologia de savana no bioma Cerrado, se fossem desconsideradas. Falsos positivos (áreas erradas classificadas como Savana Parque) podem ser refinados, *a posteriori*, com validação em campo ou dados complementares. Assim, ao concluir o experimento, chegamos à conclusão de que não existe modelo "perfeito" — a escolha depende do contexto e dos objetivos de cada projeto.

[5] A interrogação (?) na Tabela 6.4 significa que o valor da métrica não foi calculado devido à ausência de acertos.

7 CONSIDERAÇÕES FINAIS

O objetivo desta pesquisa foi centrado na busca de atributos ou padrões preditivos para ocorrência do cerrado, utilizando algoritmos de aprendizado de máquina para identificar regras associativas entre as variáveis ambientais (clima, geologia, geomorfologia, solos, declividade e água disponível no solo) na formação de tipologias vegetais.

A pesquisa explorou a aplicação de regras de associação (*Association Rules*) a partir de uma grade estatística de 1 km x 1 km, em formato vetorial, algo pioneiro. Em uma área preservada e com levantamentos temáticos integrados (RECOR), o delineamento experimental conseguiu identificar relações frequentes entre os atributos ambientais e a vegetação. Entretanto, percebeu-se que tais regras demonstram apenas coocorrência entre os atributos, sem estabelecer uma relação causal ou preditiva. Diante desta limitação, a pesquisa evoluiu para o uso de Árvores de Decisão e Random Forest, como modelos classificadores, buscando aumentar a capacidade preditiva do modelo. Apesar do risco de *overfitting*, principalmente no algoritmo J48, ambos modelos apresentaram bons resultados de acurácia global. No caso da RECOR, mesmo sem a variável clima, os resultados foram 62.58% e 62,89% para os algoritmos J48 e RF, respectivamente. Em se tratando do Bioma Cerrado, os modelos apresentaram uma acurácia global de 79,16% e 83,65%, para o J48 e Random Forest, respectivamente. Destaca-se que, no caso da RECOR, mesmo com 4 variáveis preditoras apenas, o modelo conseguiu uma predição razoável para 14 classes-alvo, sendo a classe “cerradão”, a única que os modelo não conseguiram uma boa discriminação.

Os resultados confirmaram a hipótese levantada de que algoritmos de aprendizado de máquina (*machine learning)* e técnicas de mineração de dados conseguiriam perceber os padrões associativos entre os atributos ambientais estudados. Sobre a validação, a pesquisa buscou algoritmos classificadores e os resultados apontaram que a utilização de algoritmos de modelos probabilísticos aliado a algoritmos de aprendizado supervisionado conseguiram bons relatórios de desempenhos de classificação para discriminar as classes de

savanas. Salienta-se que, enquanto a rede bayesiana elucidou a estrutura de dependências entre clima e geomorfologia, as regras de associação revelaram combinações ambientais frequentes, típicas de cada classe de cerrado, e os algoritmos J48 e Random Forest validaram, com métricas robustas, a capacidade preditiva desses padrões.

Para os dados do Bioma Cerrado, a integração com Redes Bayesianas permitiu a incorporação de inferência probabilística, resultando em um modelo híbrido capaz de melhorar a acurácia e a compreensão das relações entre os atributos ambientais e os tipos de vegetação do cerrado. O classificador BayesNet apesar de uma acurácia global modesta de 63,41%, foi o único cujo modelo conseguiu discriminar a classe minoritária da Savana Parque. A partir deste resultado, optou-se por realizar um classificador híbrido entre Random Forest e Bayes Net, a fim de que fosse possível aumentar o acerto de classificação da classe Savana Parque (classe minoritária).

Como resultado final, o classificador híbrido atingiu uma taxa de acerto em torno de 72,26%, um pouco inferior ao modelo do Random Forest, mas permitiu discriminar melhor a Savana Parque. A opção de escolha por um classificador híbrido, validado pelo valor mais alto de Recall torna-se estratégico quando o objetivo é mapear áreas de potencialidade para cerrado (savana), pois o custo ecológico de ignorar áreas reais de algum tipo de savana pode ser irreversível. Assim, um modelo híbrido foi a solução para minimizar o “trade-off”, combinando as vantagens do RF (alta precisão geral) e BN (sensibilidade a padrões raros), demonstrando a viabilidade da abordagem proposta. Entretanto, há muito espaço para melhorias no desempenho do algoritmo classificador: melhores ajustes dos hiperparâmetros, utilização de técnicas de balanceamento de classes, agrupamento de classes para diminuir a cardinalidade de algumas variáveis, entre outras. Também há espaço para inclusão de certas variáveis que não participaram deste ensaio experimental, mas que futuramente podem ser modeladas para compor o *input* de dados, como: dados de queimada, dados altimétricos, integração com informações sociais da grade estatística, entre outros dados que podem enriquecer as informações para outros tipos de estudos sobre a formação vegetal.

Através dos resultados com abordagem de aprendizado de máquina foi possível concluir que todos os atributos naturais estudados contribuíram para a modulação do cerrado em suas diversas fitofisionomias. Em especial, os atributos relacionados ao relevo e ao clima foram os que mais responderam nos modelos utilizados, demonstrando interdependência muito forte entre relevo/clima para a diversidade dos tipos se savana.

Esta pesquisa contribuiu para o avanço na descoberta de padrões em elementos ambientais, aplicáveis ao aprendizado de máquina, para inferir a probabilidade de ocorrência dos diferentes tipos de vegetação do Cerrado.

O conhecimento gerado pode apoiar a elaboração futura de mapas de potencialidade da cobertura vegetal. Desde que haja dados suficientes de levantamentos prévios dos recursos naturais de uma região de interesse, a metodologia possibilitaria inferir áreas da vegetação pristina onde o uso agrícola predomina sobre a paisagem original. Como foi dito, existem inúmeras informações geoespaciais, fornecidas por entidades públicas e privadas, que podem contribuir para aprimorar o modelo classificador utilizado nesta pesquisa.

Em um mundo impulsionado pelos novos modelos de inteligência artificial e *bigdata*, o futuro é promissor para usuários do campo de informações geoespaciais, no Brasil, com a criação da INDE e, mais recentemente, da CONGEO (Comissão Nacional de Geoinformação), onde os dados geoespaciais poderão ser intercambiáveis com mais segurança e sem grandes custos operacionais.

Em conclusão, a autora espera que os resultados deste estudo possam fornecer subsídios importantes para futuras iniciativas de conservação e restauração do Cerrado.

REFERÊNCIAS BIBLIOGRÁFICAS

AB'SABER, A. N. Os domínios morfoclimáticos na América do Sul. **Geomorfologia**, n. 52, p. 1-22, 1977. Disponível em: https://biblio.fflch.usp.br/AbSaber_AN_1348615_OsDominiosMorfoclimaticos.pdf. Acesso em: 10 fev. 2025.

AB'SABER, A. N. **Os domínios da natureza no Brasil: potencialidades paisagísticas**. São Paulo: Ateliê Editorial, 2003

ABSY, M. et al. Mise en évidence de quatre phases d'ouverture de la forêt dense dans le Sud-Est de l'Amazonie au cours des 60 000 dernières années: première comparaison avec d'autres régions tropicales. **Comptes Rendus de l'Académie de Sciences**, **Series II, Paléoclimatologie**, v. 312, p. 673-678, 1991.

AGUIAR, A. J. C. et al. Biogeography and early diversification of Tapinotaspidini oil-bees support presence of Paleocene savannas in South America. **Molecular Phylogenetics and Evolution,** 2019. Disponível em: https://doi.org/10.1016/j.ympev.2019.106692.

ARAÚJO FILHO, J. C. et al. **Avaliação, predição e mapeamento de água disponível em solos do Brasil**. Rio de Janeiro: Embrapa Solos, 2022.

ARRUDA, D. M. et al. Vegetation cover of Brazil in the last 21 ka: New insights into the Amazonian refugia and Pleistocenic arc hypotheses. **Global Ecology and Biogeography**, v.27, p.47-56, 2018.

ARRUDA, D. M.; SCHAEFER, C. E. G. R. Dinâmica climática e biogeográfica do Brasil no Último Máximo Glacial: o estado-da-arte. **Estudos Avançados USP,** n. 34, v. 98, 2020. DOI: 10.1590/s0103-4014.2020.3498.012.

BARBERI, M. **Mudanças paleoambientais na região dos cerrados do Planalto Central durante o Quaternário tardio**: **o estudo da Lagoa Bonita, DF**. 2001. 210p. Tese (Doutorado em Geologia Sedimentar) - Universidade de São Paulo, São Paulo, 2001.

BARBERI, M.; RIBEIRO, M. S. L. Evolução da paisagem nas áreas de cerrado: uma análise no tempo profundo. In: GOMES, H. (Org.) **Universo do cerrado**. Goiânia: UCG, 2008. p. 15-78.

BARBOSA, A. S. Peregrinos do Cerrado. **Revista do Museu de Arqueologia e Etnologia**, n. 5, p. 159-193, 1995.

BARBOSA, A. S. Ocupação indígena no sistema biogeográfico do cerrado. In: GOMES, H. (Org.) **Universo do Cerrado**. Goiânia: UCG, 2008. p. 79-163.

BARBOSA, C. et al. Operadores zonais em álgebra de mapas e sua aplicação a zoneamento ecológico-econômico. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 9., 1998, Santos. **Anais...** São José dos Campos: INPE, 1998. p. 487-500.

BATTARBEE, R.; W.; BINNEY H. A. (Ed.). **Natural climate variability and global warming**: **a holocene perspective**. Chichester: Wiley-Blackwell, 2008. 288 p.

BEHLING, H.; LICHTE, M. Evidence of dry and cold climatic conditions at glacial times in Tropical Southeastern Brazil. **Quaternary Research**, v. 48, p. 348-358, 1997.

BEHLING, H.; HOOGHIEMSTRA, H. Neotropical Savanna environments in space and time: late quaternary interhemispheric comparisons. In: Markgraf, V. (Ed.). **Interhemispheric climate linkages**. [S.l.]: Academic Press, 2001. p.307-323

BEHLING, H. South and southeast Brazilian grasslands during Late Quaternary times: a synthesis. **Palaeogeography, Palaeoclimatology, Palaeoecology,** v. 177, p. 19-27, 2002.

BEHLING, H.; PILLAR, V. D. Late quaternary vegetation, biodiversity and fire dynamics on the southern Brazilian highland and their implication for conservation and management of modern Araucaria forest and grassland ecosystems**. Philosophical Transactions of the Royal Society: Biological Sciences,** v.362, p.243-251, 2007.

BIGARELLA, J. J. et al. **Estrutura e origem das paisagens tropicais e subtropicais**. Florianópolis: UFSC, 1994.

BRADLEY, R. S. **Paleoclimatology: reconstructing climates of the quaternary**. 2.ed. San Diego: Academic Press, 1999. 613p.

BRASIL. MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO. **Banco de Dados e Informações Ambientais (BDiA): mapeamento de recursos naturais (MRN)**. Escala 1:250 000. Rio de Janeiro: IBGE, 2023.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5 - 32, Oct., 2001.

CASSINO, R. F.; MARTINHO, C. T; CAMINHA, S. A. F. S. A late quaternary palynological record of a palm swamp in the cerrado of central Brazil interpreted using modern analog data. **Palaeogeography, Palaeoclimatology, Palaeoecology,** 2017. Doi: 10.1016/j.palaeo.2017.08.036.

CASSINO, R. C. et al. Vegetation and fire variability in the central Cerrados (Brazil) during the Pleistocene-Holocene transition was influenced by oscillations in the SASM boundary belt. **Quaternary Science Reviews**, v. 232, e106209, 2020. Doi: doi.org/10.1016/j.quascirev.2020.106209.

CLARK, P. U. et al. The last Glacial Maximum. **Science**, v. 325, n. 710, 2009. Doi: 10.1126/science.1172873.

COLEVATTI, R. G.; LIMA, N. E.; VITORINO, L. C. The Diversification of extant angiosperms in the South America dry diagonal. In: RULL, V.; CARNAVAL, A. C. (Ed.). **Neotropical diversification: patterns and processes**. Cham: Springer Nature, 2020. p. 547-568. Disponível em: https://doi.org/10.1007/978-3-030-31167-4.

COLINVAUX, P. A. et al. A long pollen record from lowland Amazonia: forest and cooling in glacial times. **Science**, v.274, p.85-88, 1996.

COLINVAUX, P. A.; OLIVEIRA, P. E.; BUSH, M. B. Amazonian and neotropical plant communities on glacial time-scales: the failure of the aridity and refuges hypotheses. **Quaternary Science Reviews**, v. 19, p. 141-169, 2000.

COUTINHO, L. M. Aspectos ecológicos do fogo no Cerrado II: as queimadas e a dispersão de sementes em algumas espécies anemocóricas do estrato herbáceo-subarbustivo. **Boletim de Botânica**, v. 5, p. 57–63, 1977.

COUTINHO, L. M. O conceito de cerrado**. Revista Brasileira de Botânica**, v. 1, p. 17-23, 1978.

CRUZ, F. et al. Orbitally driven east–west antiphasing of South American precipitation. **Nature Geoscience**, v. 2, p. 210-14, 2009.

CULVER, S. J.; RAWSON, P. F. **Biotic response of global change**: **the last 145 million years**. Cambridge: Cambridge University Press, 2004. p. 20-34.

CUSTÓDIO, I. S. **Sistema de monção da América do Sul: padrões de variabilidade durante os últimos 21.000 anos**. 2022. Tese (Doutorado em Meteorologia) - Universidade de São Paulo, São Paulo, 2022.

DE OLIVEIRA, P. E. et al. Climate change and biogeographic connectivity across the Brazilian Cerrado. **Journal of Biogeography**, v. 1, e jbi.13732-12, 201.

EITEN, G. Vegetação do cerrado. In: PINTO, M. N. (Ed.). **Cerrado**: **caracterização, ocupação e perspectivas**. 2.ed. Brasília: UnB, SEMATEC, 1994. p. 17-73.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (EMBRAPA). Serviço Nacional de Levantamento e Conservação de Solos (Rio de Janeiro, RJ). In: REUNIÃO TÉCNICA DE LEVANTAMENTO DE SOLOS, 10., 1979, Rio de Janeiro. **Anais...** 1979. 83p. (EMBRAPA-SNLCS. Miscelânea, 1).

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; PADHRAIC, S. The KDD Process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v.39, n.11, 1996.

FIRESTONE, R. B. et al. Evidence for an extraterrestrial impact 12,900 years ago that contributed to the megafaunal extinctions and the Younger Dryas cooling. **Earth, Atmospheric, and Planetary Sciences**, v. 104, n. 41, p. 16016-16021, 2007.

GARCIA, S. et al. Clima do passado face ao presente. In: JACOBI, P. R. et al. (Org.) **Temas atuais em mudanças climáticas: para os ensinos fundamental e médio**. São Paulo: USP, 2015. p. 21-29.

HAFFER, J. Speciation in Amazonian Forest birds. **Science**, v. 165, n. 3889, p. 131-137, 1969.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. 3.ed. São Paulo: Elsevier, 2012.

HORÁK-TERRA, I. et al. Late quaternary vegetation and climate dynamics in central-eastern Brazil: insights from a ~35k cal a bp peat record in the Cerrado biome. **Journal of Quaternary Science**, v. 1, p. 1-13, 2020.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Ao completar 50 anos, Reserva Ecológica do IBGE poderá ser integrada ao Sistema Nacional de Unidades de Conservação (SNUC)**. Agência IBGE, 2025. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/42720-ao-completar-50-anos-reserva-ecologica-do-ibge-podera-ser-integrada-ao-sistema-nacional-de-unidades-de-conservacao-snuc.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Biomas e sistema costeiro-marinho do Brasil: compatível com a escala 1:250 000**. Rio de Janeiro: IBGE, 2019. 168p.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Grade estatística: nota metodológica**. Rio de Janeiro: IBGE, 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Grade estatística de dados ambientais**: **versão 2023: nota metodológica**. Rio de Janeiro: IBGE, 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Manual técnico da vegetação brasileira**: **sistema fitogeográfico; inventário das formações florestais e campestres; técnicas e manejo de coleções botânicas; procedimentos para mapeamentos**. Rio de Janeiro: IBGE, 2012. p. 15-163.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Reserva ecológica do IBGE.** Rio de Janeiro: IBGE, 2011.

JACOBS, B. F.; KINGSTON, J.D.; JACOBS, L. L. The origin of grass-dominated ecosystems. **Annals of the Missouri Botanical Garden**, v. 86, n. 2, 1999, p. 590-643.

JOUSSAUME, S. **Klima: gestern, heute, morgen**. Berlin: Springer Verlag, 1996.

KAUFFMANN, J. B.; CUMMINGS, D.L.; WARD, D.E. Relationships of fire, biomass and nutrient dynamics along a vegetation gradient in the Brazilian Cerrado. **Journal of Ecology**, v. 82, n. 3, p. 519-531, 1994.

KUPLICH, T. M.; BISPO, P. C.; VALERIANO, M. M. Análise geomorfométrica como subsídio ao mapeamento da vegetação. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 13., 2007, Florianópolis. **Anais...** São José dos Campos: INPE, 2007. p. 1607-1613.

LEDRU, M. P. et al. The last 50,000 years in the Neotropics (Southern Brazil): evolution of vegetation and climate. **Palaeogeography, Palaeoclimatology, Palaeoecology**, v. 123, n. 1/4, p. 239-257, 1996. Doi: 10.1016/0031-0182(96)00105-8.

LEDRU, M.-P. Late quaternary history and evolution of the cerrados as revealed by palynological records. In: OLIVEIRA, P. S.; MARQUIS, R. J. (Ed.). **The Cerrados of Brazil: ecology and natural history of a neotropical Savanna**. New York: Columbia University Press, 2002. p.33-50.

LEITE, P. F. **Fitogeografia do bioma Cerrado**. Brasília: IBGE, 2004.

LOIOLA, S. A. **Variações paleoclimáticas e a evolução de sistemas complexos adaptativos entre os humanos modernos**. 2014. Tese (Doutorado em Geografia) – Universidade Federal de Goiás, Goiânia, 2014.

LOWE, J; WALKER, M. **Reconstructing quaternary environments**. 3.ed. New York: Routledge, 2015. 538 p.

LUCENA, U. P.; ARAÚJO, J. C. S. Lagoa Santa, contribuição para a geografia fitobiológica e a construção acadêmica do (os) cerrado (s): discussões de Peter Lund e Eugenius Warming sobre a origem do cerrado**. Brazilian Journal of Development**, v.6, n. 11, p. 86251-86269, 2020.

MARTIN, L. et al. Astronomical forcing of contrasting rainfall changes in tropical South America between 12,400 and 8800 cal yr B.P. **Quaternary Research**, v. 47, n. 1, p. 117-122, 1997.

MASSON-DELMOTTE, V. et al. Information from paleoclimate archives. In: STOCKER, T. F. et al. (Ed.). **Climate change 2013**: **the physical science basis**. Cambridge: Cambridge University Press, 2013. p. 388-393.

McGUFFIE, K.; HENDERSON-SELLERS, A. **The climate modelling primer**. Oxford: Wiley Blackwell, 2014. p. 316-319.

MEYER, I.; WAGNER, S. The little ice age in southern South America: proxy and model based evidence. In: VIMEAUX, F.; SYLVESTRE, F.; KHODRI, M. (Ed.). **Past climate variability in South America and surrounding regions: from the last Glacial Maximum to the holocene**. Berlin: Springer, 2009. p. 395-412. DOI: 10.1007/978-90-481-2672-9_16.

NEVES, W. A.; RANGEL JUNIOR, M. J.; MURRIETA, R. S. (Orgs.). **Assim caminhou a humanidade**. São Paulo: Palas Athena, 2015. 318 p.

NIMER, E. **Climatologia do Brasil**. Rio de Janeiro: IBGE, 1979. 421p.

NOVELLO, V. F. et al. Vegetation and environmental changes in tropical South America from the last glacial to the Holocene documented by multiple cave sediment proxies. **Earth and Planetary Science Letters**, v. 524, e115717, 2019.

OLIVEIRA-FILHO, A. T.; RATTER, J. A. Vegetation physiognomies and wood flora of the Cerrado biome. In: OLIVEIRA, P. S.; MARQUIS, R. J. (Ed.). **The Cerrados of Brazil**: **ecology and natural history of a neotropical Savanna**. New York: Columbia University Press, 2002. p. 91-120

OLIVEIRA, M. J. et al. História geológica e ciência do clima: métodos e origens do estudo dos ciclos climáticos na Terra. **Terræ**, v.12, n.1, p. 3-26, 2015. Disponível em: https://www.ige.unicamp.br/terrae/V12/PDFv11/T056-1.pdf. Acesso em: 16 ago. 2020.

PEARL, J. **Probabilistic reasoning in intelligent systems**: **networks of plausible inference**. San Francisco: Morgan Kaufmann, 1988.

PEARL, J. **Causality: models, reasoning and inference**. 2.ed. Cambridge: Cambridge University Press, 2009.

PRADO, D. E.; GIBBS, P. E. Patterns of species distribuitions in the dry seasonal forests of South America. **Annals of the Missouri Botanical Garden**, v. 80, n. 4, p. 902-927, 1993.

QGIS.ORG. **QGIS: a free and open-source Geographic Information System**. 2023. Disponível em: https://www.qgis.org. Acesso em: 27 maio 2025.

QUINLAN, J. R. **C4.5: programs for machine learning**. San Francisco: Morgan Kaufmann, 1993. 302 p.

REDE CERRADO. **Defendemos o Cerrado porque ele é o nosso "berço das águas"!** Disponível em:https://redecerrado.org.br/nossa-atuacao/defesa-do-cerrado/. Acesso em 16 ago. 2021.

RIBEIRO, J. F.; WALTER, B. M. T. As principais fitofisionomias do Bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P. (Ed.). **Cerrado: ecologia e flora**. Planaltina, DF: Embrapa,1998. p. 89-166. Disponível em:

https://www.researchgate.net/publication/283072910_As_principais_fitofisionomias_do_bioma_Cerrado/link/5628c21f08ae04c2aeaeb5d9/download. Acesso em: 30 out. 2019.

ROCHA-CAMPOS, A. C.; SANTOS, P. R. A ação geológica do gelo. In: TEIXEIRA, W. et al. (Org.) **Decifrando a Terra**. São Paulo; Oficina de Textos, 2000. 568 p.

RUDDIMAN, W.F. **Earth's climate**: **past and future**. 3.ed. New York: Freeman and Company, 2014. 445 p.

SALGADO-LABOURIAU, M. L. **História ecológica da Terra**. São Paulo: Blucher, 1994. p. 255-290.

SALGADO-LABOURIAU, M. L. Alguns aspectos sobre a Paleoecologia dos Cerrados. In: SCARIOT, A.; SOUSA-SILVA, J. C.; FELFILI, J. M. (Org.). **Cerrado**: **ecologia, biodiversidade e conservação**. Brasília: MMA, 2005. p. 107-118.

SAMPAIO, G.; MARENGO, J.; NOBRE, C. A atmosfera e as mudanças, climáticas. In: BUCKERIDGE, M. S. (Org.). **Biologia & mudanças climáticas no Brasil**. São Carlos: RiMa, 2008. p.7-9.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. DOI: 10.1147/rd.33.0210.

SANT'ANNA NETO, J. L.; NERY, J. T. N. Variabilidade e mudanças climáticas no Brasil e seus impactos regionais. In: SOUZA, C.R.G. et al. (Ed.). **Quaternário do Brasil**. Ribeirão Preto: Holos, 2005. p.28-50.

SIFEDDINE, A. et al. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. **PNAS**, v. 106, n. 48, p. 20359-20364, 2009.

SIMON, M. F. et al. Evidence for adaptation to fire regimes in the tropical savannas of the Brazilian Cerrado. **International Journal of Plant Sciences**, v. 173, n. 6, p. 711-723, 2012.

STRIKIS, N. M. et al. Abrupt variations in South American monsoon rainfall during the Holocene based on a speleothem record from central-eastern Brazil. **Geology**, v. 39, p. 1075-1078, 2011.

STUTE, M. et al. Cooling of tropical Brazil (5°C) during the last Glacial Maximum. **Science**, v. 269, p. 379-383, 1995.

SUGUIO, K. **Geologia do quaternário: mudanças ambientais (passado + presente = futuro?)**. São Paulo: Paulo's, 1999. 366p.

WANG, X. et al. Interhemispheric anti-phasing of rainfall during the last glacial period. **Quaternary Science Review**, v.25, p.3391-3403, 2006.

WARMING, E. **Lagoa Santa**: **contribuição para a Geografia Phytobiologica**. Belo Horizonte, Imprensa Official do Estado de Minas Gerais, 1908. 284p.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data mining: practical machine learning tools and techniques**. 4.ed. Amsterdam: Morgan Kaufmann, 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Appendix: The WEKA workbench**. Disponível em: <https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf>. Acesso em: 5 jul. 2025.

GLOSSÁRIO DE TERMOS DE MACHINE LEARNING UTILIZADOS NA PESQUISA

Causalidade:

A **causalidade** é a relação entre causa e efeito, onde uma mudança em uma variável (causa) provoca diretamente uma mudança em outra variável (efeito). Em *Machine Learning*, inferir causalidade vai além da correlação estatística, exigindo métodos específicos para identificar e validar relações causais genuínas, frequentemente baseados em experimentos controlados (ensaios randomizados) ou técnicas de inferência causal. Na pesquisa, a abordagem e ferramenta utilizada foi a de **grafos acíclicos direcionados** (DAG's) para representar relações causais através do algoritmo BayesNet (ou rede bayesiana), pois é considerado um modelo causal estrutural. Redes Bayesianas possibilitam a modelagem de relações probabilísticas entre variáveis e quando construídas com base em conhecimento causal (ou aprendidas com técnicas apropriadas), elas podem representar relações de causas e efeitos (Pearl, 2009).

Dados transacionais:

Os dados transacionais são tipos de dados estruturados que registram eventos discretos ou interações individuais envolvendo itens. Cada registro (transação) representa um evento único (e.g., uma compra, um clique, um acesso) e contém uma lista de itens associados a esse evento. São o formato de dados primários para mineração de padrões frequentes e regras de associação. Exemplos: Carrinhos de compra em supermercado (cada carrinho é uma transação, os produtos comprados são os itens), histórico de cliques em uma sessão web, registros de diagnósticos de um paciente em uma visita. Na pesquisa, o registro (transação) considerado foi a célula de 1 km x 1km da grade estatística do IBGE. O conjunto de células (e suas respectivas variáveis ambientais integradas por célula) formou o conjunto de dados transacionais.

Padrões associativos:

Os padrões associativos são coocorrências frequentes ou relações estatísticas entre itens em um conjunto de dados. Representam regularidades observadas nos dados, como "clientes que compram o produto **A** frequentemente também compram o produto **B**". Esses padrões podem indicar correlação, mas não implicam necessariamente causalidade. Os padrões associativos são considerados a base da mineração de dados em Regras de Associação.

Validação de um Classificador:

Em *Machine Learning*, validar um classificador é o processo de avaliar rigorosamente o desempenho e a capacidade de generalização de um modelo de classificação treinado. O objetivo é estimar como o modelo se comportará ao classificar dados novos e não vistos durante o treinamento, evitando o superajuste (*overfitting*).

Como é feita a validação de um classificador?

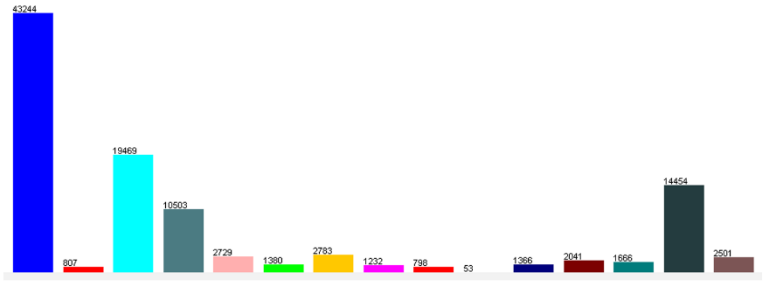
1. Divisão de Dados: Separar o conjunto de dados original em pelo menos dois subconjuntos disjuntos:
    - Conjunto de Treinamento: usado para ajustar os parâmetros do modelo.
    - Conjunto de Teste: usado apenas uma vez, no final, para avaliar o desempenho final do modelo treinado. Nunca usado durante o treinamento ou seleção de modelo.
2. Validação Cruzada: técnica onde o conjunto de treinamento é dividido em k folds (partes) de tamanho semelhante. O modelo é treinado k vezes, usando k-1 folds para treino e o fold restante para validação. A métrica de desempenho é calculada em cada iteração e depois a média das k avaliações fornece uma estimativa mais robusta do desempenho do modelo de hiperparâmetros. O conjunto de teste permanece separado.

3. Métricas de Avaliação: Calculadas no conjunto de teste ou nas iterações de validação cruzada:

- Matriz de Confusão: Tabela que mostra Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN), Falsos Negativos (FN).
- Acurácia: (VP + VN) / Total. Proporção de previsões corretas.
- Precisão: VP / (VP + FP). Proporção de previsões positivas corretas.
- Recall ou teste de Sensibilidade: VP / (VP + FN). Proporção de casos positivos reais identificados.
- F1-Score: Média harmônica entre Precisão e Recall (2 * (Precisão * Recall) / (Precisão + Recall)).
- Área Sob a Curva ROC (AUC-ROC): Mede a capacidade do modelo de distinguir entre classes, considerando todas as taxas de falsos positivos e verdadeiros positivos.

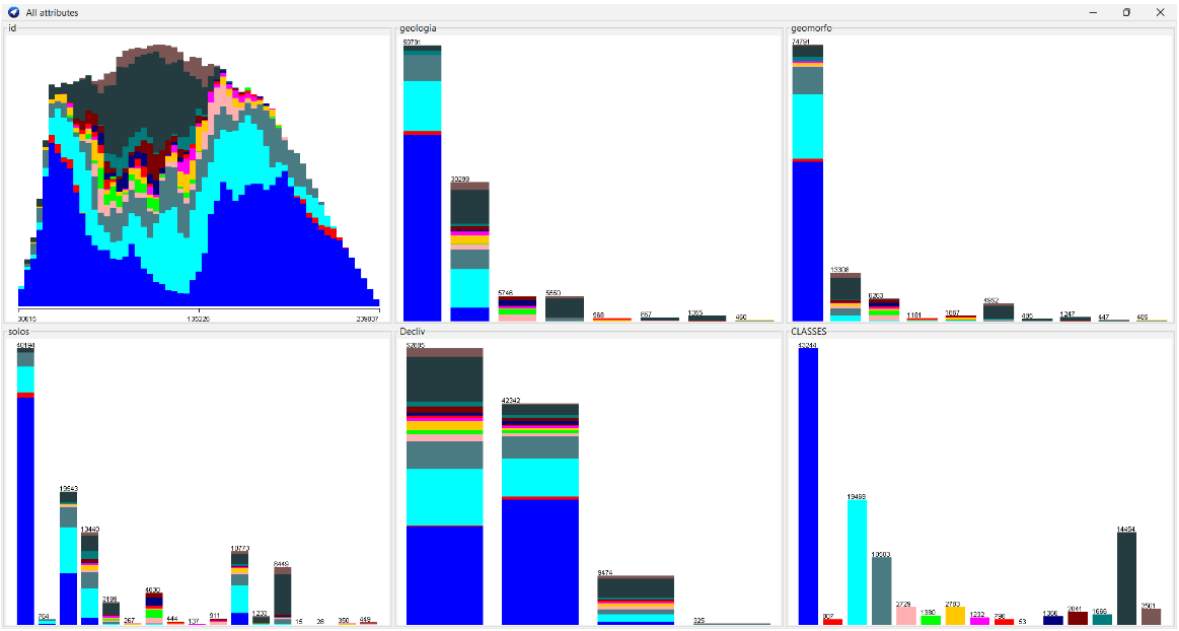
APÊNDICE A – VARIÁVEIS-ALVO E PREDITORAS DA RECOR

VARIÁVEIS-ALVO DA VEGETAÇÃO DE CERRADO



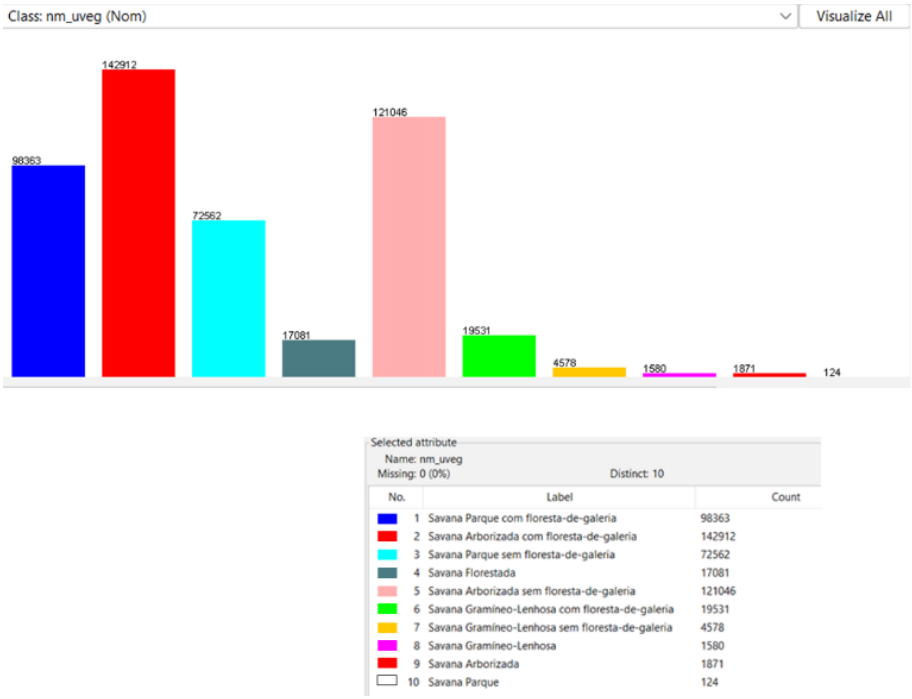
No.	Label	Count	Weight
1	Cerrado	43244	43244
2	Cerradão	807	807
3	Campo Cerrado	19469	19469
4	Campo Sujo	10503	10503
5	Capoeira	2729	2729
6	Floresta Galeria Paludosa	1380	1380
7	Campo Úmido com Murundum	2783	2783
8	Floresta Galeria Seca	1232	1232
9	Brejo	798	798
10	Brejo com Murundum	53	53
11	Vereda	1366	1366
12	Campo Úmido	2041	2041
13	Campo Sujo com Murundum	1666	1666
14	Campo Limpo	14454	14454
15	Campo Limpo com Murundum	2501	2501

VARIÁVEIS PREDITORAS - DISPLAY DO WEKA

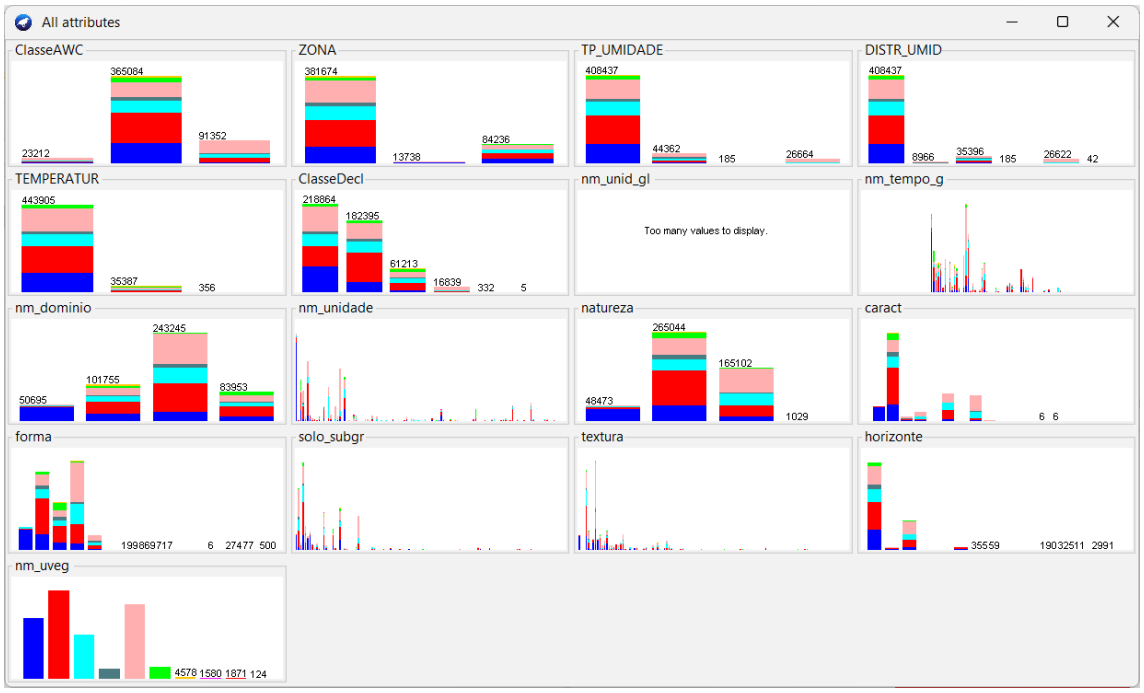


APÊNDICE B – VARIÁVEIS-ALVO E PREDITORAS DO BIOMA CERRADO

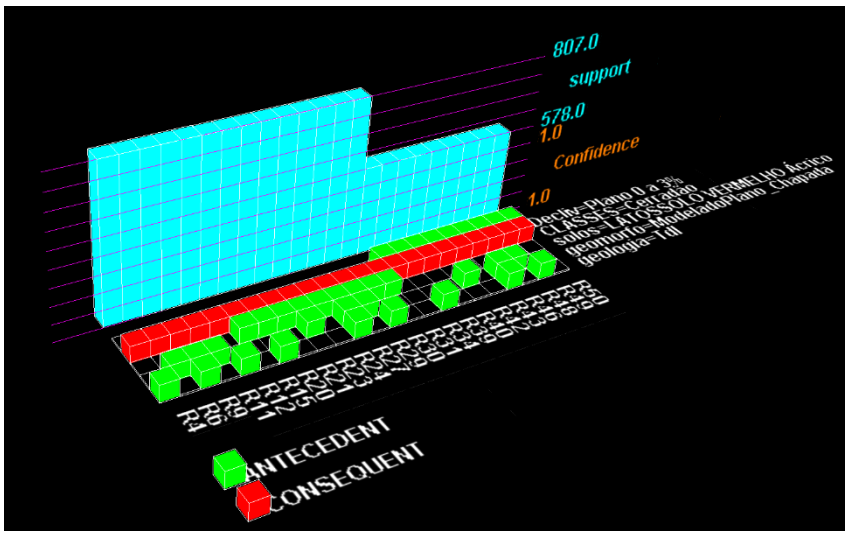
VARIÁVEIS-ALVO

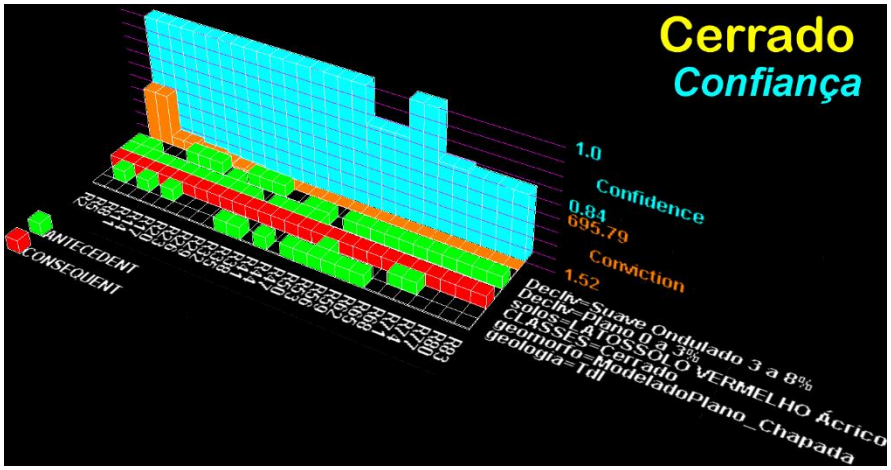


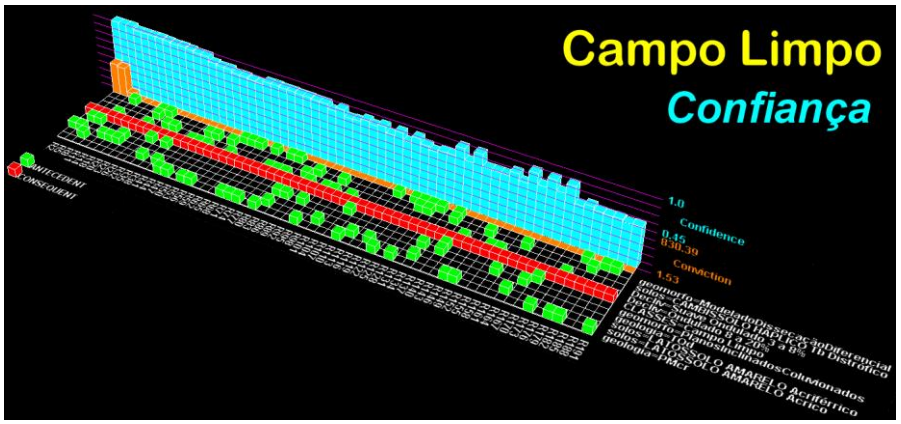
VARIÁVEIS PREDITORAS NO DISPLAY DO WEKA



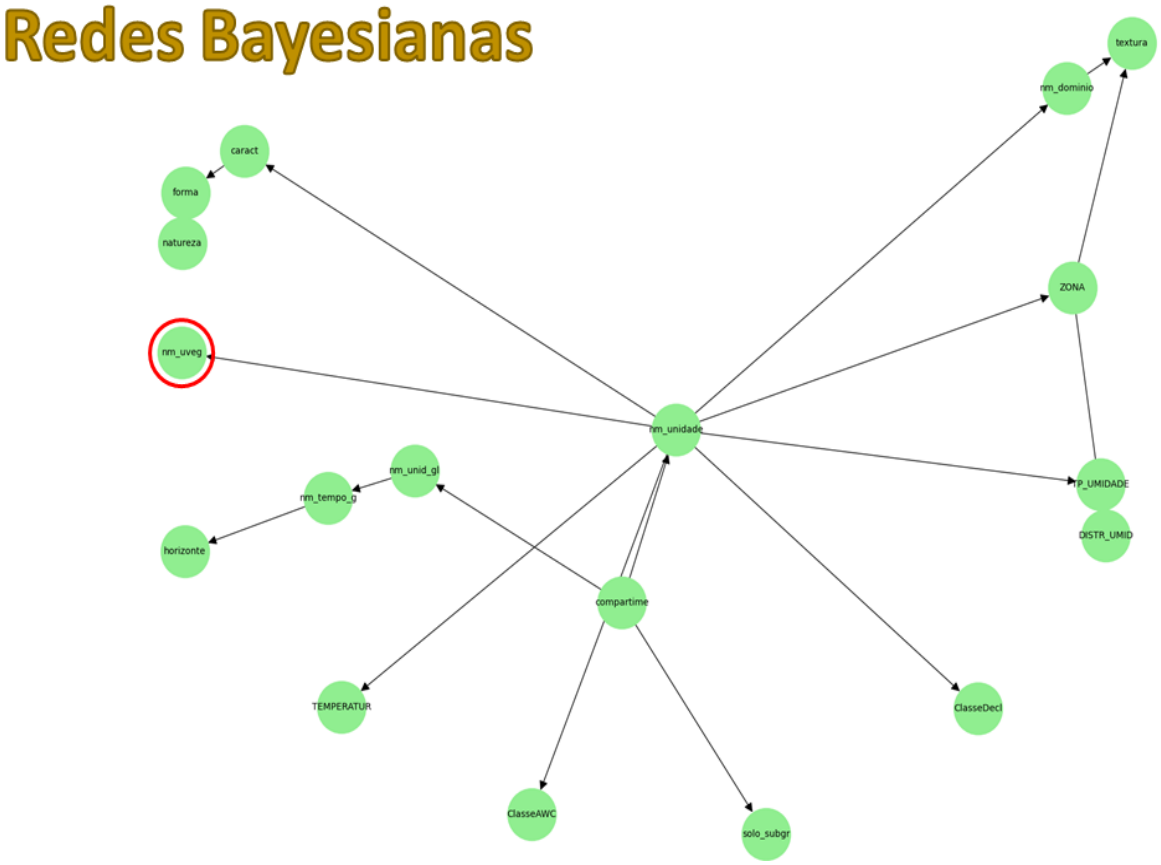
APÊNDICE C – EXEMPLOS DE VISUALIZAÇÂO DAS REGRAS DE ASSOCIAÇÃO NO WEKA







APÊNDICE D – GRAFO DA REDE BAYESIANA COM TODAS AS VARIÁVEIS AMBIENTAIS



A variável-alvo está circulada em vermelho.

APÊNDICE E – EXEMPLOS DE TABELAS DE DISTRIBUIÇÃO DE PROBABILIDADE (TDP) GERADAS PELO CLASSIFICADOR BAYES NET

Resultados para AWC e Declividade para Savana Arborizada.

Distribuição Condicional de ClasseAWC:

Para nm_uveg = Savana Arborizada:

ClasseAWC	phi (ClasseAWC)
ClasseAWC(Classe AD4)	0.0329
ClasseAWC(Classe AD5)	0.9556
ClasseAWC(Classe AD6)	0.0116

Distribuição Condicional de ClasseDecl:

Para nm_uveg = Savana Arborizada:

ClasseDecl	phi(ClasseDecl)
ClasseDecl(Forte Ondulado 20 a 45%)	0.0165
ClasseDecl(Montanhoso 45 a 75%)	0.0005
ClasseDecl(Ondulado 8 a 20%)	0.3365
ClasseDecl(Plano - 0 a 3%)	0.1179
ClasseDecl(Suave Ondulado 3 a 8%)	0.5285

A pesquisa gerou Tabelas de Distribuição de Probabilidade (TDP) para todas as classes das variáveis ambientais envolvidas, nas 10 tipologias de Savana estudadas. As tabelas demonstradas acima apenas demonstram uma pequena parte da saída do algoritmo BayesNet para a Savana Arborizada somente relacionadas às variáveis AWC (disponibilidade de água no solo) e declividade.

APÊNDICE F - EXEMPLO DE REGRAS DE ASSOCIAÇÃO GERADAS NO WEKA

=== Run information ===

Scheme: weka.associations.Apriori -N 200 -T 0 -C 0.9 -D 0.1 -U 1.0 -M 0.6 -S 0.05 -A -c -1

Relation: df_savana_atualizado_pedo-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.instance.SubsetByExpression-EATT18 is "Savana Arborizada"

Instances: 1871

Attributes: 18

ClasseAWC

ZONA

TP_UMIDADE

DISTR_UMID

TEMPERATUR

ClasseDecl

nm_unid_gl

nm_tempo_g

nm_dominio

nm_unidade

natureza

caract

forma

compartime

solo_subgr

textura

horizonte

nm_uveg

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.6 (1123 instances)

Minimum metric <confidence>: 0.9

Significance level: 0.05

Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 4

Size of set of large itemsets L(4): 1

Best rules found:

1. ZONA=Tropical Brasil Central 1871 ==> nm_uveg=Savana Arborizada 1871 conf:(1)

2. ClasseAWC=Classe AD5 1800 ==> nm_uveg=Savana Arborizada 1800 conf:(1)

3. ClasseAWC=Classe AD5 ZONA=Tropical Brasil Central 1800 ==> nm_uveg=Savana Arborizada 1800 conf:(1)

4. TP_UMIDADE=semi-úmido 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

5. DISTR_UMID=4 a 5 meses secos 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

6. ZONA=Tropical Brasil Central TP_UMIDADE=semi-úmido 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

7. ZONA=Tropical Brasil Central DISTR_UMID=4 a 5 meses secos 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

8. TP_UMIDADE=semi-úmido DISTR_UMID=4 a 5 meses secos 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

9. ZONA=Tropical Brasil Central TP_UMIDADE=semi-úmido DISTR_UMID=4 a 5 meses secos 1662 ==> nm_uveg=Savana Arborizada 1662 conf:(1)

10. ClasseAWC=Classe AD5 TP_UMIDADE=semi-úmido 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

11. ClasseAWC=Classe AD5 DISTR_UMID=4 a 5 meses secos 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

12. ClasseAWC=Classe AD5 ZONA=Tropical Brasil Central TP_UMIDADE=semi-úmido 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

13. ClasseAWC=Classe AD5 ZONA=Tropical Brasil Central DISTR_UMID=4 a 5 meses secos 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

14. ClasseAWC=Classe AD5 TP_UMIDADE=semi-úmido DISTR_UMID=4 a 5 meses secos 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

15. ClasseAWC=Classe AD5 ZONA=Tropical Brasil Central TP_UMIDADE=semi-úmido DISTR_UMID=4 a 5 meses secos 1596 ==> nm_uveg=Savana Arborizada 1596 conf:(1)

16. horizonte=A moderado 1138 ==> nm_uveg=Savana Arborizada 1138 conf:(1)

17. ZONA=Tropical Brasil Central horizonte=A moderado 1138 ==> nm_uveg=Savana Arborizada 1138 conf:(1)

APÊNDICE G – CÓDIGOS PYTHON UTILIZADOS NA PESQUISA

CLASSIFICADOR DE REDES BAYESIANAS

```
import pandas as pd
from pgmpy.models import BayesianNetwork
from pgmpy.estimators import HillClimbSearch, BicScore, BayesianEstimator
from pgmpy.inference import VariableElimination
import networkx as nx
import matplotlib.pyplot as plt

# Carregar os dados
df = pd.read_csv(r"D:/TeseResultados/TestesAmostragens/amostra_crf.csv", sep=';', encoding='ISO-8859-1')

# Definição dos grupos de variáveis
grupos = {
    "Grupo_1": ['ZONA', 'TP_UMIDADE', 'DISTR_UMID', 'TEMPERATUR', 'nm_uveg'],
    "Grupo_2": ['nm_dominio', 'nm_unidade', 'natureza', 'caract', 'forma', 'compartime', 'nm_uveg'],
    "Grupo_3": ['solo_subgr', 'textura', 'horizonte', 'nm_uveg'],
    "Grupo_4": ['ClasseAWC', 'ClasseDecl', 'nm_uveg']
}

# Dicionário para armazenar os modelos e inferências
resultados = {}

for nome_grupo, variaveis in grupos.items():
```

```
    print(f"Rodando Bayesian Network para {nome_grupo}...")

    # Filtrar as variáveis do grupo
    df_subset = df[variaveis].dropna()

    # Estimar a estrutura da rede
    est_hc = HillClimbSearch(df_subset)
    melhor_modelo = est_hc.estimate(scoring_method=BicScore(df_subset))

    # Criar modelo Bayesian Network
    modelo = BayesianNetwork(melhor_modelo.edges())
    modelo.fit(df_subset, estimator=BayesianEstimator, prior_type="BDeu")

    # Inferência
    infer = VariableElimination(modelo)
    resultados[nome_grupo] = (modelo, infer)

    # Gerar gráfico da rede
    plt.figure(figsize=(10, 6))
    G = nx.DiGraph()
    G.add_edges_from(modelo.edges())
    print("Nós do grafo:", G.nodes())
    print("Arestas do grafo:", G.edges())
    pos = nx.spring_layout(G)  # Sem seed

    nx.draw(G, pos, with_labels=True, node_color='lightblue', edge_color='gray',
node_size=3000, font_size=10)
```

```
    plt.title(f"Estrutura da Rede Bayesiana - {nome_grupo}")
    plt.savefig(f"D:/ALDA/Downloads/Bayesian_Network_{nome_grupo}.png")
    plt.show()

    # Tabela de probabilidades para cada variável em relação a nm_uveg
    for var in variaveis:
        if var != 'nm_uveg':
            print(f"Distribuição Condicional de {var} dado nm_uveg no {nome_grupo}:")
            print(infer.query([var], evidence={'nm_uveg': 'Savana Arborizada'}))
            print("-" * 80)

print("Processamento concluído!")
```

APÊNDICE H – CÓDIGO PYTHON DAS REDES BAYESIANAS, TABELAS E GRÁFICOS DE PROBABILIDADE

#Trecho 2: Rede Bayesiana Integrada com as tabelas dos gráficos de barra-

```
import os
import re
import matplotlib.pyplot as plt
import numpy as np

# Diretórios
resultados_dir = "D:/ALDA/Downloads/Resultados_Bayes"
bar_chart_group_dir = os.path.join(resultados_dir, "BarCharts_GroupSavanna")
if not os.path.exists(bar_chart_group_dir):
    os.makedirs(bar_chart_group_dir)

# Função para parse do arquivo de distribuições
def parse_distribuicoes(file_path):
    distrib_dict = {}
    with open(file_path, "r", encoding="utf-8") as f:
        lines = f.readlines()

    current_var = None
    current_condition = None
    table_lines = []
    re_line = re.compile(r'^\s*\|\s*(.*?)\s*\|\s*(.*?)\s*\|')

    for line in lines:
        line = line.strip()
        if line.startswith("Distribuição Condicional de"):
            m = re.search(r"Distribuição Condicional de (.+):", line)
            if m:
                current_var = m.group(1).strip()
                distrib_dict[current_var] = {}
        elif line.startswith("Para nm_uveg ="):
```

```
            if current_var and current_condition and table_lines:
                parsed_table = {}
                for tbl_line in table_lines:
                    match = re_line.match(tbl_line)
                    if match:
                        cat, prob = match.group(1).strip(), match.group(2).strip()
                        try:
                            parsed_table[cat] = float(prob)
                        except:
                            continue
                distrib_dict[current_var][current_condition] = parsed_table
                table_lines = []
            m = re.search(r"Para nm_uveg = (.+):", line)
            if m:
                current_condition = m.group(1).strip()
        elif line.startswith("----------------"):
            if current_var and current_condition and table_lines:
                parsed_table = {}
                for tbl_line in table_lines:
                    match = re_line.match(tbl_line)
                    if match:
                        cat, prob = match.group(1).strip(), match.group(2).strip()
                        try:
                            parsed_table[cat] = float(prob)
                        except:
                            continue
                distrib_dict[current_var][current_condition] = parsed_table
                table_lines = []
        else:
            if line.startswith("|"):
                table_lines.append(line)
    if current_var and current_condition and table_lines:
        parsed_table = {}
        for tbl_line in table_lines:
            match = re_line.match(tbl_line)
```

```
            if match:
                cat, prob = match.group(1).strip(), match.group(2).strip()
                try:
                    parsed_table[cat] = float(prob)
                except:
                    continue
        distrib_dict[current_var][current_condition] = parsed_table
    return distrib_dict

# Lista de tipos de savana
savanna_types = [
    "Savana Arborizada",
    "Savana Arborizada com floresta-de-galeria",
    "Savana Arborizada sem floresta-de-galeria",
    "Savana Florestada",
    "Savana Gramíneo-Lenhosa",
    "Savana Gramíneo-Lenhosa com floresta-de-galeria",
    "Savana Gramíneo-Lenhosa sem floresta-de-galeria",
    "Savana Parque",
    "Savana Parque com floresta-de-galeria",
    "Savana Parque sem floresta-de-galeria"
]

# Função para gerar gráficos e tabelas
def plot_savanna_class_probabilities(all_group_data, savanna_types, threshold=0.6):
    cores_variaveis = {
        'ZONA': 'red',
        'TP_UMIDADE': 'red',
        'DISTR_UMID': 'red',
        'TEMPERATUR': 'red',
        'nm_unid_gl': 'green',
        'nm_tempo_g': 'green',
        'nm_dominio': 'blue',
        'nm_unidade': 'blue',
        'natureza': 'blue',
```

```
    'caract': 'blue',
    'forma': 'blue',
    'compartime': 'blue',
    'solo_subgr': 'purple',
    'textura': 'purple',
    'horizonte': 'purple',
    'ClasseAWC': 'orange',
    'ClasseDecl': 'orange'
}

for savanna_type in savanna_types:
    consolidated = {}
    for group_name, group_data in all_group_data.items():
        for var, cond_dict in group_data.items():
            if savanna_type in cond_dict:
                for cat, prob in cond_dict[savanna_type].items():
                    if prob > threshold:
                        if cat not in consolidated or prob > consolidated[cat][1]:
                            consolidated[cat] = (var, prob)

    if not consolidated:
        print(f"Sem registros para {savanna_type} com prob > {threshold}.")
        continue

    records = [(v[0], k, v[1]) for k, v in consolidated.items()]
    records.sort(key=lambda x: x[2], reverse=True)

    # Salvar tabela
    savanna_filename = re.sub(r'\W+', '_', savanna_type)
    table_filename = f"Tabela_{savanna_filename}.txt"
    table_path = os.path.join(bar_chart_group_dir, table_filename)

    with open(table_path, "w", encoding="utf-8") as f:
        f.write("Variável\tCategoria\tProbabilidade\n")
        for entry in records:
```

```
        f.write(f"{entry[0]}\t{entry[1]}\t{entry[2]:.4f}\n")
print(f"Tabela salva: {table_path}")

# Gerar gráfico
colors = [cores_variaveis.get(entry[0], 'gray') for entry in records]
labels = [entry[1] for entry in records]
values = [entry[2] for entry in records]

x = np.arange(len(labels))
plt.figure(figsize=(max(12, len(labels)*0.5), 8))
bars = plt.bar(x, values, color=colors)

# Legenda
unique_vars = list(set([entry[0] for entry in records]))
legend_elements = [
    plt.Line2D([0], [0], marker='s', color='w', label=var,
               markerfacecolor=cores_variaveis.get(var, 'gray'), markersize=10)
    for var in unique_vars
]
plt.legend(handles=legend_elements, loc='lower left', fontsize=8)

plt.xticks(x, labels, rotation=90, ha="right", fontsize=10)
plt.ylabel("Probabilidade")
plt.title(f"Probabilidades Consolidadas - {savanna_type} (> {threshold*100}%)")

for bar, val in zip(bars, values):
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, height, f"{val:.2f}",
             ha="center", va="bottom", fontsize=7)

plt.tight_layout()
graph_filename = f"Grafico_{savanna_filename}.png"
graph_path = os.path.join(bar_chart_group_dir, graph_filename)
plt.savefig(graph_path)
plt.close()
```

```
        print(f"Gráfico salvo: {graph_path}")

# Processamento principal
all_group_data = {}
for file_name in os.listdir(resultados_dir):
    if file_name.startswith("Distribuicoes_") and file_name.endswith(".txt"):
        group_file = os.path.join(resultados_dir, file_name)
        group_name = file_name.replace("Distribuicoes_", "").replace(".txt", "")
        print(f"Processando grupo {group_name}...")
        group_data = parse_distribuicoes(group_file)
        all_group_data[group_name] = group_data

# Gerar gráficos e tabelas
plot_savanna_class_probabilities(all_group_data, savanna_types, threshold=0.6)

print("Processo concluído com sucesso!")
```

APÊNDICE I – CÓDIGO PYTHON PARA ANÁLISE MCA

```
import os
import pandas as pd
import prince
import matplotlib.pyplot as plt
import numpy as np


# Caminho para o arquivo CSV
file_path = input("Digite o caminho completo do arquivo CSV gerado: ")
output_folder = os.path.join(os.path.dirname(file_path), 'MCA_graficos')
os.makedirs(output_folder, exist_ok=True)

# Carregar os dados com amostragem de 10%
try:
    df = pd.read_csv(file_path, encoding='latin1', sep=';')
    print("Dados carregados com sucesso!")
    df = df.sample(frac=0.1, random_state=42)  # Amostragem de 10%
    print(f"Amostra reduzida para {len(df)} instâncias.")
    if 'ind_gre' in df.columns:
        df.drop(columns=['ind_gre'], inplace=True)  # Remover coluna ind_gre
        print("Coluna 'ind_gre' removida com sucesso!")
except Exception as e:
    print(f"Erro ao carregar os dados: {e}")
    exit()

# Definir as variáveis, excluindo a coluna de ID, se existir
variaveis = [col for col in df.columns if col != 'indice_gre']
df_features = df[variaveis]

# Converter variáveis categóricas
df_features = df_features.apply(lambda x: x.astype('category') if x.dtypes ==
'object' else x)

# Realizar MCA
try:
    mca = prince.MCA(n_components=2)
    mca_result = mca.fit_transform(df_features)
    print("MCA realizada com sucesso!")
except Exception as e:
    print(f"Erro ao realizar a MCA: {e}")
    exit()
```

```
# Criar gráfico MCA
fig, ax = plt.subplots(figsize=(14, 12))
colors = plt.cm.get_cmap('tab20', len(df_features.columns))

# Plotar variáveis preditoras
for i, col in enumerate(df_features.columns):
    ax.text(mca_result.iloc[i, 0], mca_result.iloc[i, 1], col, fontsize=12,
color=colors(i), weight='bold', ha='center',
            bbox=dict(facecolor='white', edgecolor='none', alpha=0.7))

# Adicionar subtipos de savana
if 'nm_uveg' in df.columns:
    subtipos = df['nm_uveg'].unique()
    for subtipo in subtipos:
        subtipo_indices = df['nm_uveg'] == subtipo
        ax.scatter(
            mca_result.loc[subtipo_indices, 0],
            mca_result.loc[subtipo_indices, 1],
            label=subtipo,
            alpha=0.8,
            s=80,
            edgecolors='black'
        )

# Configurações do gráfico
ax.legend(title="Subtipos de Savana (nm_uveg)", bbox_to_anchor=(1.05, 1),
loc='upper left')
ax.set_title('MCA das Amostras com Variável Alvo (nm_uveg)', fontsize=16)
ax.set_xlabel("Componente MCA 1", fontsize=12)
ax.set_ylabel("Componente MCA 2", fontsize=12)
ax.grid(alpha=0.3)
plt.tight_layout()
# Salvar e exibir o gráfico
output_path = os.path.join(output_folder, 'MCA_com_nm_uveg.png')
plt.savefig(output_path, dpi=300)
print(f"Gráfico salvo em {output_path}")
plt.show()
```

ANEXO A – CARTA CRONOESTRATIGRÁFICA INTERNACIONAL

