Cássio Augusto Rodrigues Bettim

Identificação de marcadores genéticos biogeograficamente informativos para amostras do tráfico brasileiro de *Cannabis sativa*: Uma abordagem de aprendizado de máquina para rastreamento forense de origem

Porto Alegre

Março de 2025

Cássio Augusto Rodrigues Bettim

Identificação de marcadores genéticos biogeograficamente informativos para amostras do tráfico brasileiro de *Cannabis sativa*: Uma abordagem de aprendizado de máquina para rastreamento forense de origem

Dissertação apresentada ao Programa de Pós graduação em Biologia Celular e Molecular da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do título de Mestre em Biologia Celular e Molecular.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre

Março de 2025

Agradecimentos

Gostaria de expressar minha gratidão àqueles que sempre estiveram ao meu lado ao longo destes dois anos de jornada acadêmica.

Aos meus pais, Cintia Oliveira e Alisson Bettim, e ao meu padrasto, Júlio Lírio, pelo amparo constante e apoio incondicional, pelo suporte emocional e financeiro, e por sempre me incentivarem a ser resiliente e confiar em minhas próprias capacidades. Toda conquista minha também é de vocês. Amo-lhes infinitamente e lhes sou eternamente grato.

Ao meu orientador, Márcio, pela oportunidade e pelo acolhimento, pela compreensão nos momentos de difíceis, pela confiança depositada em meu trabalho e pelo apoio em todas as formas que importam.

Aos meus colegas do *Structural Bioinformatics and Computational Biology Lab*, pelo aprendizado, histórias e risadas partilhadas, essenciais para tornar o dia a dia da pesquisa muito mais leve. Mais do que colegas, hoje lhes considero grandes amigos.

À todos os meus amigos, aos de longa convivência até os recém chegados, pelos momentos de descontração, pelas palavras de incentivo, por escutarem meus desabafos, por serem alicerce em momentos incerteza, por sempre estarem ao meu lado quando precisei. Junto de vocês, a vida se torna muito mais bela de ser vivida.

Resumo

Cannabis sativa é o psicoativo mais distribuído globalmente e a droga ilícita mais consumida na América Latina e no Brasil, com mais de 800 toneladas da droga apreendidas no ano de 2024. Além disso, os esforços dedicados da Polícia Federal (PF) para combater o tráfico de drogas e erradicar grandes cultivares em território nacional, principalmente na região conhecida como "Polígono da Maconha", levaram ao aumento de práticas como o cultivo *indoor* em menor escala e formas alternativas de tráfico, como o envio de sementes por serviços postais, vindas de países como Colombia, Uruguai, Paraguai e do continente Europeu. Nesse contexto, há uma necessidade crescente de metodologias que auxiliem na tarefa de rastreabilidade de plantas, sementes ou subprodutos de *C. sativa* traficados, diferenciando-os e predizendo fenótipos de interesse como sua origem de cultivo, o que é crucial para a inteligência policial. No presente trabalho, buscamos avaliar diferentes protocolos de extração de variantes do tipo *SNP* e microssatélites (*SSR*), marcadores amplamente utilizados na área forense, associados à estratégias de seleção de atributos para redução de dimensionalidade e algoritmos clássicos de aprendizado supervisionado, verificando sua aplicabilidade para construção de classificadores para predição de origem de cultivo. Diferentes métricas para avaliar modelos de classificação foram consideradas, além da execução de Análises de Coordenadas Principais (PCoA) e agrupamento hierárquico como análises complementares. Para tal, foram utilizados dados de sequenciamento completo de 38 amostras de sementes e subprodutos de *C. sativa* apreendidas pela PF, vindas de quatro rotas de tráfico conhecidas: Paraguai, Colombia, Polígono da Maconha e um grupo estrangeiro, identificado em relatórios policiais como possivelmente Europeias. Em nossos experimentos, verificamos que, de modo geral, os algoritmos *Support Vector Classifiers (SVC)* associados a diferentes estratégias de seleção de atributos apresentaram desempenho superior na predição da origem das amostras, sendo capazes de classificar corretamente, em tarefas distintas, todas as amostras usando dois subconjuntos distintos de 9 *SSRs* e 5 *SNPs* localizados em diferentes cromossomos, os quais também foram capazes de distinguir visualmente os quatro grupos com base em sua origem biogeográfica em PCoA e agrupamento hierárquico. Nosso estudo identificou com sucesso marcadores informativos do tipo *SNP* e *SSR* capazes de distinguir amostras de *C.*

sativa relacionadas a importantes rotas de tráfico brasileiras, utilizando de abordagens clássicas de aprendizado de máquina supervisionado, de forma inédita na literatura científica atual. No entanto, estudos adicionais em novos e maiores grupos de amostras são necessários para validar sua aplicação em condições experimentais reais.

Abstract

Cannabis sativa is the most widely distributed psychoactive globally and the most consumed illicit drug in Latin America and Brazil, with more than 800 tons of the drug seized in 2024. In addition, the dedicated efforts of the Brazilian Federal Police (BFP) to combat drug trafficking and eradicate large cultivars on national territory, especially in the region known as the “Marijuana Polygon”, have led to an increase in practices such as indoor cultivation on a smaller scale and alternative forms of trafficking, such as sending seeds by postal service from countries such as Colombia, Uruguay, Paraguay and the European continent. In this context, there is a growing need for methodologies that assist in the task of tracing trafficked *C. sativa* plants, seeds or by-products, differentiating them and predicting phenotypes of interest such as their origin of cultivation, which is crucial for police intelligence. In this study, we sought to evaluate different protocols for extracting SNP and microsatellite (SSR) variants, markers widely used in forensics, associated with feature selection strategies to reduce dimensionality and classic supervised learning algorithms, verifying their applicability for building classifiers to predict the origin of *C. sativa* samples. Different metrics for evaluating classification models were considered, as well as Principal Coordinate Analysis (PCoA) and hierarchical clustering as complementary analyses. To do this, we used whole genome sequencing data from 38 samples of *C. sativa* seeds and by-products seized by the BFP, coming from four known trafficking routes: Paraguay, Colombia, the Marijuana Polygon and a foreign group, identified in police reports as possibly European. In our experiments, we found that, in general, the Support Vector Classifiers (SVC) algorithms associated with different attribute selection strategies showed superior performance in predicting the origin of the samples, being able to correctly classify, in different tasks, all the samples using two distinct subsets of 9 SSRs and 5 SNPs located in different chromosomes, which were also able to visually distinguish the four groups based on their biogeographic origin in PCoA and hierarchical clustering. In summary, our study has successfully identified informative markers capable of distinguishing *C. sativa* samples related to important Brazilian trafficking routes, using classical supervised machine learning approaches, in a way that is unprecedented in current

scientific literature. However, further studies on new and larger groups of samples are needed to validate its application in real experimental conditions.

Lista de Figuras

Figure1 – Representaçãoilustrativadasdiferentesmorfologiasdassubspéciesde*C.sativa* 15

Figure2 – Representaçãoilustrativadasfloresmachoefêmeade*C.sativa*eseusprincipais

18 25 Figure4 – Representaçãográficadosistemade*BAGGing*emumalgoritmo*RandomForest*. 27

Figure3 – Representaçãográficadotreinamentoemumalgoritmodeaprendizado. 29 Figure6 – Representaçãográficadeumhiperplanodemargemrígidaemumespaçobidi-

supervisionado.

30 Figure7 – Representaçãográficadoaumentodedimensionalidadeutilizando*kerneltrick*.. 32 Figure8 – Tiposdesubstituiçõeserazãodetransiçõesetransversõesobservadasnoconjunto

84 Figure9 – Avaliação dos Modelos de classificação baseados em SVC utilizando *kernel* linear 86 Figure 10 – Avaliação dos Modelos de classificação baseados em GB . . . . . 87 Figure 11 – Avaliação dos Modelos de classificação baseados em RF . . . . . 88 Figure 12 – Matrizes de desequilíbrio de ligação (LD) entre os 5 SNPs selecionados pelo

modelo *SVC* através dos diferentes grupos de origem 89

Figure 13 – *Principal Coordinate Analysis (PCoA)* baseada na distância de Manhattan entre as amostras..... 92

Figure 14 – *Hierarchical Clustering* baseado no método *UPGMA* e distância de Manhattan . 93

Lista de Tabelas

Table1 – Parâmetrosforensescalculadosparaos5SNPsselecionados. 9
Table2 – Frequênciasalélicasdos5SNPsselecionados . . . . . Table3 –. 0
*Fixationindex(FST)*entreosquatrogruposdeorigem 9
0
9
1

Lista de abreviaturas e siglas

AEC	Antes da Era Comum
THC	*Delta*-9-tetrahidrocanabidiol
CBD	Canabidiol
OLA	Ácido olivético
CBGA	Ácido Canabigerólico
TCHA	Ácido Tetrahidrocanabidiólico
CBDA	Ácido Canabidiólico
CB1	Receptor Canabinóide tipo 1
CB2	Receptor Canabinóide tipo 2
MP	*Marijuana Polygon*
PG	*Paraguay*
CO	*Colombia*
FG	*Foreign group*
UNODC	*United Nations Office on Drugs and Crime*
ANVISA	Agência Nacional de Vigilância Sanitária
PF	Polícia Federal
BFP	*Brazilian Federal Police*
PRF	Polícia Rodoviária Federal

SSRs	*Short Sequence Repeats*
SNPs	*Single Nucleotide Polymorphisms*
ML	*Machine Learning*
FS	*FeatureSelection*
RF	*Random Forest*
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
GB	*Gradient Boosting*
LOO	*Leave-One-Out*
PCOA	Análise de Coordenadas Principais
CHR	Cromossomo
POS	Posição
LD	Desequilíbiro de ligação

Índice

CAPÍTULO 1: Introdução **13**

1. ***Cannabis sativa*: aspectos históricos, morfológicos, bioquímicos, legais e implicações forenses** **13**

1.1 Origem e expansão da *C. sativa* 13

1.2 Sobre a espécie e suas diferentes classificações 14

1.3 A biossíntese dos Canabinóides 17

1.4 Consumo e caráter legal ao redor do mundo 19

1.5 Tráfico em território nacional e Métodos forenses de detecção 20

1.6 Visão geral . . 23

2. **Conceitos básicos de aprendizado de máquina** **24**

2.1 *Random Forest* 25

2.2 *Gradient Boosting* 27

2.3 *Support Vector Classifier* 29

2.3.1 Margem rígida 30

2.3.1 Margens suaves . 31

2.3.2 Função *kernel* 32

2.3.2 *SVMs* em problemas multiclasse 33

2.3.3 Otimização de hiperparâmetros 33

3 *GridSearch* . 34

2.3.4 *Random Search* 34

4 Otimização Bayesiana 34

2.4 Validação cruzada 35

2.4.1 *Hold-out validation* 35

2.4.2 *K-fold* 35

Leave-One-Out . . .

2

2.4.

3

2.6 Seleção de Atributos 36
2.6.1 Métodos *Filter* 36
2.6.2 Métodos *Wrapper* 37
2.6.3 Métodos *Embedded* 37
2.7 Visão geral 38

Capítulo 2: Objetivos **39**

Capítulo 3: Apresentação do artigo científico **41**

Capítulo 4: Predição de origem a partir de variantes do tipo *SNP* utilizando aprendizado de máquina supervisionado **79**
1. Materiais e métodos **79**
2. Resultados adicionais e Discussão **84**

Capítulo 5: Conclusão geral **94**

Referências **96**

CAPÍTULO 1: Introdução

1. *Cannabis sativa*: aspectos históricos, morfológicos, bioquímicos, legais e implicações forenses

1.1 Origem e expansão da *C. sativa*

Cannabis sativa (*C. sativa*) é considerada uma das plantas mais antigas a serem cultivadas pela humanidade para diferentes propósitos, utilizadas para produção de papel e fibras, como fonte de alimento, fins medicinais e farmacêuticos, além do uso em rituais religiosos devido a suas propriedades psicoativas. Sua distribuição ao redor do globo e o surgimento de variedades são fortemente associados à formação de assentamentos humanos e antigas rotas comerciais utilizadas, tornando difícil a tarefa de determinar seu centro de origem, mas é comum atribuí-lo às regiões da Ásia central e oriental (CLARKE; MERLIN, 2013; MERLIN, 1972).

Registros fósseis encontrados no Japão e datados de aproximadamente 8000 aEC indicam o consumo do fruto (semente) de *C. sativa* como fonte nutricional (HUSSAIN et al., 2021; MC-PARTLAND; HEGMAN; LONG, 2019; PISANTI; BIFULCO, 2018), sendo uma fonte rica em proteínas, lipídios e ácidos graxos, principalmente para povos da Ásia central e oriental (IFTIKHAR et al., 2021; LAPIERRE; MONTHONY; TORKAMANEH, 2023). *Shennong Bencaojing*, também entitulado de *Classic of the Materia Medica*, foi talvez um dos primeiros registros escritos do uso medicinal da *C. sativa*, e compreende uma compilação em texto de tradições orais registradas durante a dinastia Han (206 aEC - 220 EC), mas cuja autoria é atribuída a *Shennong* (2700 aEC), figura mitológica da cultura chinesa tida como pai da medicina tradicional (CROCQ, 2020; PISANTI; BIFULCO, 2018; XIE et al., 2023). Também, no antigo Egito, o uso medicinal de *C. sativa* para tratamento de inflamações e doenças oftamlmológicas pode ser observado em documentos históricos como papiros médicos de Ramesseum (1700 aEC) e papiros de Ebers (1550 aEC) (XIE et al., 2023; CROCQ, 2020).

Quanto a sua difusão dentro do continente Europeu e no Mediterrâneo, uma das teorias é de que sua forma já domesticada tenha sido trazida da Ásia central pelos Citas, povo de pastores nômades equestres da antiguidade clássica, entre os séculos VIII e III aEC durante sua expansão pela estepe pôntico-cáspia (PISANTI; BIFULCO, 2018). No entanto, achados de pólen de *C. sativa* através do território Europeu reforçam a teoria alternativa de que a planta teria se espalhado da Asia para a Europa ainda em sua forma selvagem durante o período Pleistoceno, sendo domesticada durante a Era do Cobre e do Bronze (RULL, 2022; MCPARTLAND; GUY; HEGMAN, 2018). Já no território do grande continente Africano, acredita-se que tenha sido disseminada pela ação de comerciantes conectados com a India e pela expansão do império Árabe durante o século XIII EC, sendo utilizada para tratamentos de mordidas de serpentes, analgesia durante partos, auxiliando na sintomatologia de malária, febre, asma, entre outras efermidades (NETO; PIERRO; FERNANDES, 2023; ZUARDI, 2006).

Finalmente, o aparecimento da *C. sativa* nas Américas está relacionado ao processo de colonização europeia durante o século XVI EC. Trabalhos recentes sugerem que a produção no Novo Mundo era estimulada pelas diferentes nações colonizadoras, principalmente para produção de tecidos, cordas e velas marítimas (RULL, 2022), enquanto o cultivo das variantes com efeito psicoativo mantinha-se restrito ao Velho Continente, sendo apenas introduzida anos mais tarde, especialmente na América Latina, através de trabalhadores vindos das Indias após a abolição da escravatura. Por outro lado, apesar da falta de registros documentais, também há trabalhos que defendem a ideia de que a *C. sativa* foi trazida às Américas do Sul e Central através do tráfico de escravos Angolanos - devido ao fato de muitos dos termos sinônimos usados no Brasil para referir-se à planta são de origem Angolana -, estando presente na formação dos primeiros quilombos (ASSUNÇAO, 1995; NETO; PIERRO; FERNANDES, 2023; ZUARDI, 2006; CARLINI, 2006).

1.2 Sobre a espécie e suas diferentes classificações

C. sativa é um organismo dióico, pertencente ao gênero *Cannabis* e à família *Cannabaceae*, cujo genoma está organizado em 10 pares cromossômicos, sendo 9 autossômicos e um par sexual - XX para fêmeas e XY para machos. Seu primeiro exemplar foi descrito e catalogado em 1753 por *Carl Linnaeus* em sua obra "*Species Plantarum*" como *C. sativa L.* Mais tarde, em 1785, Jean-Baptiste

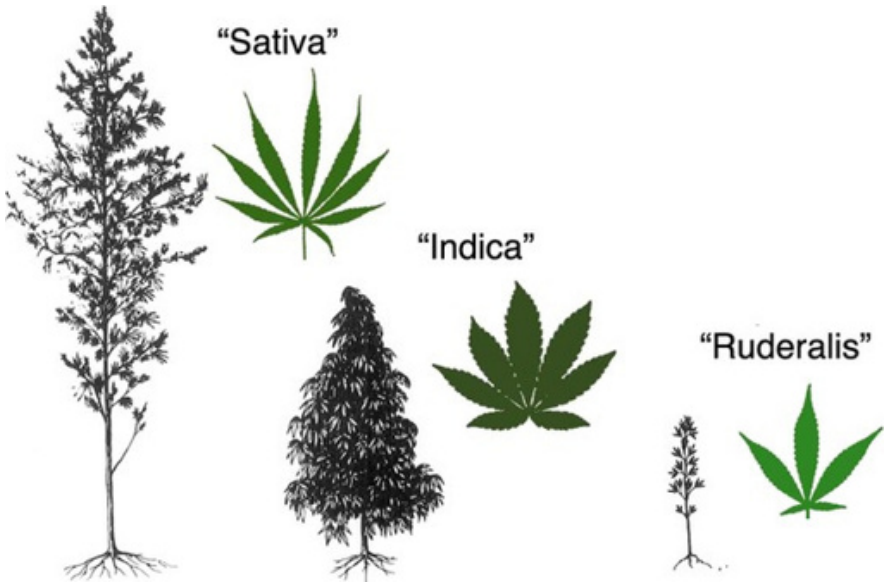


Figura 1 – **Representação ilustrativa das diferentes morfologias das subspécies de *C. sativa***.
Fonte: Adaptado de Anderson (1980).

Lamarck descreve o que viria ser considerado uma segunda espécie do gênero a partir de exemplares coletados na Índia, também presentes no Afeganistão e Paquistão, batizada assim de *C. indica Lam.*, e que diferia da *C. sativa* em diversos aspectos, não apenas morfológicos mas também de odor forte e presença de efeitos psicoativos (WATTS, 2006). Apenas entre 1922 e 1924, surgiria uma possível nova espécie identificada na região sudeste da Rússia, primeiramente nomeada por Nikolai Vavilov como *C. sativa var. spontanea* - e mais tarde conhecida por seu sinônimo *C. ruderalis* (termo do latim botânico que significa "erva daninha"), cunhado por D.E. Janischewsky -, referente a espécimes considerados selvagens com características morfológicas diferentes das demais, como plantas mais baixas e galhos mais encurtados (JANISCHEVSKY, 1924; VAVILOV, 1922). Sua classificação taxonômica é considerada até hoje um tópico de discussão entre botânicos, havendo diferenças tanto dentro da nomenclatura científica quanto entre esta e a taxonomia vernacular, adotada pela cultura e linguagem popular, principalmente no nível de espécie. No entanto, a classificação mais aceita atualmente é a de que *C. sativa* trata-se de uma espécie monotípica, cujas variedades devem ser organizadas na hierarquia em subespécies, sendo assim: *C. sativa subsp. sativa, C. sativa subsp. indica* e *C. sativa subsp. ruderalis* (SMALL; CRONQUIST, 1976; MCPARTLAND; GUY, 2017; MCPARTLAND, 2018).

A primeira forma de diferenciação entre as subsespécies, como já brevemente mencionado,

pode ser feita de acordo com suas diferenças morfológicas (Figura 1). *C. sativa L.* é mais variável e são relativamente mais altas, entre 1,5 até aproximadamente 5 metros de altura, galhos mais espalhados, tricomas esparsos e folhas palminérvias com 5-7 folíolos serrilhados de formato longo e lanceolado. *C. indica* são plantas mais curtas, de 0,5 a 1,5 metros e formato "piramidal", galhos e tricomas mais densos, caule menos fibroso e folhas palminérvias com 7-11 folíolos serrilhados mais largos e oblanceolados. *C. ruderalis* são consideravelmente mais baixas, atingindo no máximo 0,5 metros, normalmente sem galhos e folhas palminérvias mais elípticas com 3-5 folíolos (ANDERSON, 1980; MUROVEC et al., 2022).

Uma segunda forma de diferencição, e possivelmente a mais utilizada, pode ocorrer com relação a seu propósito primário, que se dá pela concentração de Tetrahidrocanabidiol (*delta*-9-THC, ou apenas THC) e Canabidiol (CBD) (ACOSTA; ALMIRALL, 2021; MUROVEC et al., 2022; MALABADI et al., 2023) - os principais compostos canabinóides produzidos pela planta e que serão melhor abordados posteriormente -, separando-as nas variedades como cânhamo e maconha. Neste cenário, utiliza-se o termo cânhamo, ou *hemp*, para referir-se a *C. sativa L.* cuja porcentagem de THC (responsável pela resposta psicoativa) é inferior a 0,3% do peso seco e, geralmente, maiores concentrações de CBD, apresentando uma proporção CBD/THC abaixo de 1, sendo cultivadas para fins industriais, principalmente para: fabricação de fibra de cânhamo destinada a indústria têxtil e em produtos como cordas, vestuário e papel; produção de sementes destinadas ao setor alimentício devido a seu alto valor nutricional; e extração de óleo para produção de biodiesel. Já a maconha, também conhecida por *weed*,*Ganja*,*Marijuana*, entre outros, faz referência a variantes cuja concentração de THC é superior a 0,3%, podendo alcançar até 30% em algumas cepas comerciais, com uma proporção CBD/THC acima de 1. Esses limites foram propostos pela primeira vez em 1976 por Small e Cronquist (1976) como um critério distintivo entre as variedades comerciais do cânhamo e a *C. sativa* de uso recreativo ou medicinal. Desde então, diversos países ao redor do mundo, incluindo Brasil, Estados Unidos (EUA), China, Canadá, França e Bélgica, continuam a adotá-los como referência para regulamentação legal (RAIHAN; BIJOY, 2023).

Isso nos leva a terceira forma de classificação, relativa aos diferentes quimiotipos de *C. sativa*, que origina-se da utilização da proporção entre THC e CBD. Plantas com altos teores de THC (relação CBD/THC entre 0,00 e 0,05) são classificadas como Tipo 1 ou THC-dominantes; aquelas com valores intermediários (relação CBD/THC entre 0,05 e 3,00) são denominadas Tipo 2

ou balanceadas; enquanto aquelas com baixos teores de THC (relação CBD/THC entre 15,00 e 25,00) são classificadas como Tipo 3 ou CBD-dominantes (MUROVEC et al., 2022). Essa categorização é viabilizada pelo fato de que, por serem produtos de um mesmo precursor biossintético, a relação CBD/THC mantém-se praticamente constante ao longo de todo o ciclo de desenvolvimento da planta (STAGINNUS; ZöRNTLEIN; MEIJER, 2014; PACIFICO et al., 2007).

1.3 A biossíntese dos Canabinóides

O termo canabinóide refere-se, de forma genérica, a um grupo de compostos químicos que podem ser encontrados em plantas (fitocanabinóides), como a *C. sativa*, ou em organismos mamíferos (endocanabinóides), e que interagem com receptores do sistema canabinóide (MEAH et al., 2021). A *C. sativa* apresenta um extenso perfil de mais de 550 compostos já identificados, como terpenos, flavonóides, alcalóides, mas são os fitocanabinóides os de maior relevância. (HUSSAIN et al., 2021; ROCK; PARKER, 2020; RADWAN et al., 2017). Existem mais de uma centena de fitocanabinóides já reportados, sendo os mais abundantes e de maior importância o THC e o CBD, e suplementados por outros como canabicromeno (CBC), canabigerol (CBG), canabinol (CBN), entre outros (TAHIR et al., 2021).

O THC e CBD são sintetizados em estruturas chamadas tricomas gladulares encontrados em grande quantidade envolvendo os pistilos presentes nas inflorescências da *C. sativa* fêmea (PUNJA; SUTTON; KIM, 2023) (Figura 2), mas também em menores concentrações em outras parte da planta como caule, folhas e as prórprias flores (United Nations Office on Drugs and Crime, 2022). Flores de plantas macho possuem consideravelmente menor quantidade de tricomas em suas flores e, logo, baixíssima concentração de fitocanabinóides. A biossíntese do THC e CBD pode ser brevemente explicada da seguinte forma (TAHIR et al., 2021; DEGENHARDT; STEHLE; KAYSER, 2017):
a enzima tetracetídeo sintase (*TKS*) cataliza a condensação sequencial do Hexanoil-CoA com três moléculas de Malonil-CoA gerando 3,5,7-Trioxododecanoil-CoA, que por sua vez é ciclizado e aromatizado , com perda da coenzima A, pela Ácido olivetólico ciclase (*OAC*) em Ácido olivético (*OLA*). Um grupamento prenil é adicionado ao OLA por uma preniltransferase aromática, dando origem ao Ácido Canabigerólico (*CBGA*), considerada a molécula "mãe" dos fitocanabinóides. A partir do *CBGA*, ocorre a transformação para os precursores dos principais canabinóides de interesse,

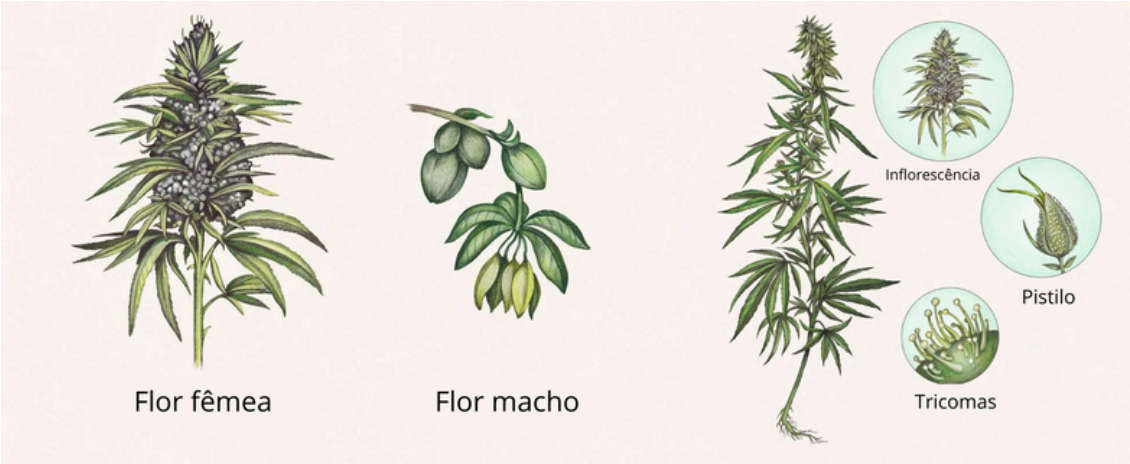


Figura 2 – **Representação ilustrativa das flores macho e fêmea de *C. sativa* e seus principais componentes**. Fonte: Adaptado do *website* Ontario Cannabis Store, disponível em: < https://ocs.ca/blogs/cannabis- basics/cannabis- plant- anatomy- the- par ts- of- the- plant>

o ácido *delta*-9-tetrahidrocanabidiólico (THCA) e o ácido canabidiólico (CBDA), pela ação das enzimas THCA sintase e CBDA sintase, respectivamente. Cerca de 95% do THC e CBD presentes na *C. sativa*, assim como outros fitocanabinóides, estão em sua forma ácida - THCA e CBDA -, que só são convertidos em seus produtos finais através de reações não enzimáticas, que podem ocorrer durante o armazenamento do produto ou, principalmente, após o aquecimento (TAHIR et al., 2021

Dentro do organismo, o THC e o CBD ligam-se nos receptores canabinóides tipo 1 (CB1) e tipo 2 (CB2), uma classe de receptores acoplados à proteína G. O CB1 está presente principalmente no cérebro de mamíferos, nas regiões do cortex frontal, hipocampo, cerebelo, medula espinhal e sistema nervoso periférico, e também já identificado no sistema cardiovascular e trato gastrointestinal (ARAUJO; ALMEIDA; ARAUJO, 2023; CHAYASIRISOBHON, 2021), fortemente associado à inibição da liberação de neurotransmissores glutamato e ácido gama-aminobutírico (GABA) (SHAHBAZI et al., 2020), enquanto o CB2 é mais abundante em células do sistema imune e hematopoiéticas, porém também presente em outros tecidos (DEVINSKY et al., 2014; ARAUJO; ALMEIDA; ARAUJO, 2023), sendo associado a funções de neuroproteção, manutenção da massa óssea e redução de processos inflamatórios (SHAHBAZI et al., 2020). A ligação do THC - principal fitocanabinóide entorpecente produzido pela planta - aos receptores, como resultado das mais variadas formas de consumo da planta para fins recreativos, acarreta em alterações somáticas (torpor, tontura, sensação de "flutuação", taquicardia, fraqueza, etc), cognitivas (instrospecção, sonolência, perda de concentração, descoordenação motora, etc.) e de percepção (desfoque, perda do senso de tempo,

distorções de visão, fadiga, entre outras)(HOCH et al., 2024; STELLA, 2023). O CBD por sua vez, considerado o componente não-psicoativo, é de interesse medicinal, sendo utilizado para melhora sintomática de casos de epilepsia infantil, dor crônica, ansiedade, sintomas relacionados ao câncer, além de neutralizar efeitos negativos do THC com relação à memória, humor e cognição (BILBAO; SPANAGEL, 2022; HOCH et al., 2024; RAPIN et al., 2021).

1.4 Consumo e caráter legal ao redor do mundo

A *C. sativa* e seus subprodutos, quando utilizada para fins recreativos, são, ao mesmo tempo, os psicoativos e drogas ilícitas mais consumidos e distribuídos em todo o mundo (United Nations Office on Drugs and Crime (UNODC), 2024), totalizando algo próximo a 147 milhões de usuários, cerca de 2,5% da população mundial (World Health Organization, 2025). Os dados apontam porcentagens particularmente altas nos EUA, onde no ano de 2019, 16,9% dos adultos americanos consumiam *C. sativa* (maconha) para fins recreativos e 45% relatou já ter utilizado pelo menos alguma vez na vida (FARRELLY et al., 2023). No Canadá, no mesmo ano, 21% da população relatou o uso da *C. sativa*, incluindo fins medicinais e não-medicinais (Health Canada, 2019). Já na América do Sul, o destaque vai para o Chile e Uruguai, cuja prevalência de uso nos grupos de adolescentes e adultos é de 14,5% e 9,3%, respectivamente, seguidos por Argentina (7,8%) e Costa Rica (4,8%), também no ano de 2019 (CASTILLO-CARNIGLIA et al., 2020). No Brasil, o consumo da *C. sativa* como psicoativo variou de 3,4 a 3,1% entre os anos de 2012 e 2015, segundo diferentes relatórios, sendo a substância ilícita mais utilizada no país (BASTOS et al., 2017; Ministério da Justiça e Segurança Pública, 2021; RANSING et al., 2021).

Apesar da *C. sativa* ser considerada ilegal na grande maioria dos países, nota-se uma tendência das políticas em torno de seu consumo tem se tornado cada vez mais liberalizadas ao redor do mundo. Considerando os dados mais recentes sobre o tema, os países que hoje consideram o uso da *C. sativa* legal tanto para fins médicos quanto para fins recreativos são África do Sul, Alemanha, Canadá, Chile, Georgia, Luxemburgo, Tailândia, Uruguai, além de 24 estados, 3 territórios e o distrito da Columbia nos EUA e o Território da Capital Australiana. Em outros países, o uso medicinal é permitido por lei mas o uso recreativo varia entre ilegal ou descriminalizado até certa quantidade - por descriminalização, entende-se que o ato continua sendo ilegal porém não é passível de acarretar

restrição de liberdade. Como exemplo, podemos citar Portugal, Espanha, França, Itália, Suíça e República Tcheca no continente Europeu; os territórios Norte e Sul Australianos; Marrocos no continente Africano; Argentina, Chile, Colombia, Peru, Equador na América Latina e determinados estados Norte Americanos como Louisiana, Ohio e Mississipi.

No Brasil, desde 2015, a Agência Nacional de Vigilância Sanitária (ANVISA) já regulamenta e importação de medicamentos à base de CBD, incluindo-os no rol de medicamentos controlados (Agência Nacional de Vigilância Sanitária (ANVISA), 2015), e em 2019 estabeleceu os requisitos para a comercialização de produtos de *C. sativa* para fins medicinais no país (Agência Nacional de Vigilância Sanitária (ANVISA), 2019). Quanto ao uso recreativo, em junho de 2024, o Supremo Tribunal Federal do Brasil decidiu por maioria também adotar a política de descriminalização, definindo um limite de até 40 gramas de *C. sativa* (maconha) ou até seis plantas fêmeas como parâmetros de diferenciação do usuário comum e traficantes, tornado a posse para consumo pessoal apenas um ato ilícito administrativo. A decisão surge como uma forma de estabelecer critérios mais objetivos - mas que também são circunstanciais - para o combate ao tráfico, em contraste a falta dos mesmos e ao caráter subjetivo da Lei nº 11.343/2006, que institui a Sistema Nacional de Políticas Públicas sobre Drogas, que implicava em práticas como seletividade penal e encarceramento em massa de grupos sociais mais vulneráveis (SILVA; ROSA, 2020).

1.5 Tráfico em território nacional e Métodos forenses de detecção

Em concordância com o quadro global, a *C. sativa* (maconha) é a droga ilícita mais consumida em todo território brasileiro (Ministério da Justiça e Segurança Pública, 2021). Estima-se que 70% do mercado ilegal interno seja abastecido pelo tráfico internacional, vindo de áreas de cultivo espalhadas principalmente pela fronteira Brasil-Paraguai, especialmente no estado de Mato Grosso do Sul - suprindo as macrorregiões Sul, Sudeste e Centro-Oeste do país -, enquanto os 30% restantes sejam oriundos do cultivo em território nacional, especialmente da região conhecida como "Polígono da Maconha", composta por 13 cidades do Sertão nordestino e porções do médio e sub-médio do Rio São Fransciso, abrangendo os estados de Pernambuco e Bahia - e suprindo assim as macrorregiões Nordeste e Norte (FRAGA, 2019; FETT et al., 2018; FETT et al., 2020; RIBEIRO et al., 2020). De fato, no ano de 2024, 808 toneladas de maconha foram apreendidas em rodovias federais pela Polícia

Rodoviária Federal (PRF), 23,23% a mais do que o ano anterior, com destaque para os estados do Paraná (276,71 toneladas), Mato Grosso do Sul (257,83 toneladas) e Bahia (95,11 toneladas) e as rodovias BR-163, BR-267 e BR-277 (Ministério das Comunicações, 2025). Além disso, operações da Polícia Federal (PF) voltadas para a erradicação de grandes platios de *C. sativa* superaram a marca de 600 toneladas no mesmo período com mais de 260 plantios destruídos, principalmente no estados de Pernambuco, Bahia, Maranhão e Pará (Polícia Federal, 2024; G1, 2025). Em adição à estas rotas, devido ao esforço de órgãos como a PF e PRF em erradicar plantios em grande escala da planta, o envio de sementes por serviços postais - uma forma alternativa de tráfico já observada no tráfico internacional - vem sendo adotada e aumentado nos últimos anos (G1, 2024; Polícia Federal, 2022; FETT et al., 2021), visando o cultivo *indoor* e em menor escala, o que torna mais difícil de serem localizados por órgãos policiais e possibilita o cultivo de cepas com maior concentração de THC pelo controle mais rigoroso do ambiente. Estas sementes são enviadas não só de países vizinhos, como Uruguai, pioneiro na América Latina a legalizar o consumo de maconha no ano de 2013, e Colômbia, que recentemente ganharam destaque como novas portas de entrada para o tráfico (ASSIS et al., 2025), como também do continente Europeu, um importante fornecedor do tráfico inter naiconal.

Em 2022, a *United Nations Office on Drugs and Crime (UNODC)* publica o manual intitulado "Métodos recomendados para identificação e análise de Cannabis e produtos derivados", visando reunir uma série de sugestões de metodologias analíticas já validadas para identificação de diferentes tipos de amostras de *C. sativa*. De forma breve, as metodologias adotadas se classificam em: métodos físicos, métodos químicos e métodos baseados em DNA (United Nations Office on Drugs and Crime, 2022). Os métodos físicos baseiam-se na examinação das características morfológicas de nível macro, como estrutura foliar e tamanho das flores, e microscópico, como a morfologia dos tricomas. Os métodos químicos, por sua vez, baseiam-se na verificação da presença ou medição da concentração de canabinóides em diferentes partes da planta, e podem variar desde testes colorimétricos simples como o *Fast Blue BB* (amplamente utilizado pela PF), até técnicas mais avançadas como cromatografia em camada fina (*TLC*), cromatografia gasosa - espectrometria de massas (*GC-MS*) e/ou cromatografia líquida - espectrometria de massas (*LC-MS*), a depender do requerimento analítico do teste (se é qualitativo ou quantitativo).

Por fim, métodos baseados em DNA tem emergido com o avanço e maior aceessibilidade de

custo das tencologias relacionadas à biologia molecular, e podem ser uma excelente alternativa em casos onde a morfologia da planta não esteja preservada e não seja possível fazer a distinção de traços, ou então para identificação de amostras extremamente degradadas com baixas concentrações de canabinóides. Além disso, a utilização de marcadores genéticos possibilita não apenas a identificação, mas também individualização (isto é, estabelecer uma ligação entre uma amostra individual e sua fonte) e rastreabilidade das amostras, características muito relevantes no contexto investigativo forense. Dentre os marcadores genéticos explorados para aplicação forense na identificação de *C. sativa*, destacam-se as *Short Tandem Repeats* (*STRs*), também chamadas de *Short Sequence Repeats* (*SSRs*) ou microssatélites (RIBEIRO et al., 2020; FETT et al., 2018; XIA et al., 2022; ONOFRI; MANDOLINO, 2017 e os *Single Nucleotide Polimorphisms (SNPs)* (BARBARIć; BEZBRADICA, 2023; NUNZIO; BARROT-FEIXAT; GANGITANO, 2024; ONOFRI; MANDOLINO, 2017).

SSRs são repetições em tandem de 2-6 pares de base distribuídos de forma variada em diferentes regiões do genoma, especialmente em regiões não codificantes como introns, *UTRs* (do inglês *Untranslated Regions*) e espaços intergênicos, mas também presentes em exons (AB-DURAKHMONOV, 2016). São marcadores amplamente utilizados para fins de identificação e rastreabilidade de amostras oriundas de cenas de crime por sua natureza altamente polimórfica, sendo considerados o padrão ouro na genética forense contemporânea (FILHO et al., 2020; KOWALCZYK et al., 2018) sendo a base do *CODIS (Combined DNA Index System)*, banco de dados fundado pelo *FBI (Federal Bureau of Investigation)* e aplicado para identificação humana, além de terem demonstrado potencial como marcadores para rastreabilidade de amostras de *C. sativa* em trabalhos recentes (RIBEIRO et al., 2020; FETT et al., 2018). Os SNPs, por sua vez, são mutações pontuais de uma única base, que podem ser classificadas como transições, quando bases de uma classe são substituídas por bases da mesma classe (purinas por purinas ou pirimidinas por pirimidinas), ou como transversões, quando bases de uma classe são substituídas pela outra classe (purinas por pirimidinas, ou vice-versa) (GUO et al., 2017). *SNPs* são, em sua maioria, bi-alélicos, porém também podem ser tri ou tetra-alélicos (GAO; LI; ZHAN, 2018). Essa característica lhes confere menor variabilidade entre indivíduos e um caráter menos informativo como marcadores quando comparados aos SSRs, de natureza multialélica (FILHO et al., 2020). Apesar disso, algumas das vantagens da utilização dos *SNPs* são sua praticidade de serem amplificados em reações *multiplex PCR* (do inglês, *Polymerase Chain Reaction*) e capacidade de serem caracterizados a partir de pequenos

amplicons, comuns em amostras degradadas ou em quantidades diminutas, além de possuírem uma taxa de mutação mais baixa se comparada a marcadores SSR (FILHO et al., 2020). Não obstante, estudos recentes obtiveram sucesso na diferenciação de múltiplos cultivares de *C. sativa* através de painéis selecionados de marcadores *SNP* e técnicas de clusterização (NUNZIO; BARROT-FEIXAT; GANGITANO, 2023; NUNZIO; BARROT-FEIXAT; GANGITANO, 2024; JIN et al., 2021; CULL; JOLY, 2025), tanto para fins forenses quanto comerciais.

1.6 Visão geral

Na primeira seção deste capítulo, concluímos que a relação da *C. sativa* com a espécie humana é milenar, com registros históricos e arqueológicos indicando seu uso para alimentação, medicina e rituais desde o Japão neolítico até o Egito Antigo, e que sua disseminação global está fortemente ligada às migrações humanas, ao comércio e processos de colonização. Também vimos que, taxonomicamente, trata-se de uma espécie monotípica e com variedades no nível de subespécie, com diferentes formas de diferenciação a depender do contexto empregado, sendo a principal delas a concentração de THC, que a divide entre as variedades maconha e cânhamo. Além disso, tratamos da biossíntese do THC e CBD que ocorre na estrutura dos tricomas da *C. sativa* fêmea, sendo estes canabinóides de maior interesse na planta, e que ligam-se aos receptores CB1 e CB2 no organismo humano. O THC é responsável pelos efeitos psicoativos associados ao uso recreativo, enquanto o CBD está relacionado ao uso terapêutico, sendo utilizado no tratamento de diversas condições médicas.

Foram abordados também o consumo e caráter legal da *C. sativa* em diversos países ao redor do mundo, apresentando a tendência global de adotar políticas de legalização tanto do consumo recreativo quando para fins medicinais, com ênfase no Brasil onde o consumo recreativo foi recentemente descrminalizado. Apresentamos também o atual contexto do tráfico nacional de *C. sativa*, esta que é a droga ilícita mais consumida no país, que é abastecido, majoritariamente, pela fronteira com o Paraguai e pela região do Polígono da Maconha no Nordeste brasileiro, mas também por outros países latinoamericanos como Uruguai e Colômbia, assim como pelo continente Europeu. Por fim, abordamos as metodologias recomendadas para identificação forense de *C. sativa*, com destque aos métodos baseados em DNA, destacando os marcadores do tipo *SNP* e *SSR* como os mais

bem explorados neste contexto.

Considerando as características mencionadas e o potencial desses dois tipos de marcadores moleculares para aplicações forenses, especialmente no que tange a rastreabilidade, este trabalho explora protocolos distintos para a utilização de *SNPs* e *SSRs* como variáveis informativas na determinação da origem de cultivo de amostras de *C. sativa* interceptadas por órgãos policiais na rede de tráfico no Brasil, associando-os à estratégias clássicas de aprendizado de máquina. Na próxima seção deste capítulo, serão explicados conceitos básicos de aprendizado de máquina, assim como as estratégias centrais utilizadas neste trabalho a partir de dados genéticos destes mesmos marcadores para realização desta tarefa.

2. Conceitos básicos de aprendizado de máquina

Aprendizado de máquina (*ML*, do inglês *Machine learning*) refere-se a criação de modelos preditivos (combinação de algoritmos, parâmetros e hiperparâmetros ajustáveis) que buscam relações lógico-matemáticas entre váriáveis *x* e um desfecho *y* a partir da observação de um conjunto de dados e reconhecimento de padrões existentes (LORENA; CARVALHO, 2007; GREENER et al., 2021). Aurélien Geron, autor do livro "Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow", define *ML* como "a ciência (e a arte) de programar computadores de modo que eles possam aprender a partir de dados". Nesse sentido, a aplicação de técnicas de *ML* torna-se especialmente eficaz em dados de alta dimensionalidade e complexidade cuja a capacidade humana não conseguiria identificar tais padrões com a mesma facilidade e rapidez, o que é um aspecto bastante característico de grandes *datasets* biológicos, principalmente tratando-se de estudos ômicos (GREENER et al., 2021; XU; JACKSON, 2019). De modo geral, podemos dividir o *ML* em dois grandes grupos: aprendizado não-supervisionado e aprendizado supervisionado. No aprendizado não-supervisionado, os dados não possuem um rótulo pré-estabelecido e o principal objetivo é agrupá-los utilizando alguma métrica de qualidade, formando o que chamamos de *clusters*, ou seja, trata-se de uma análise exploratória e de reconhecimento para identificar o que cada grupo formado significa no contexto do problema a ser investigado. Por outro lado, no aprendizado supervisionado, todas as amostras são catalogadas e possuem um rótulo, também chamado de classe, atribuído previamente

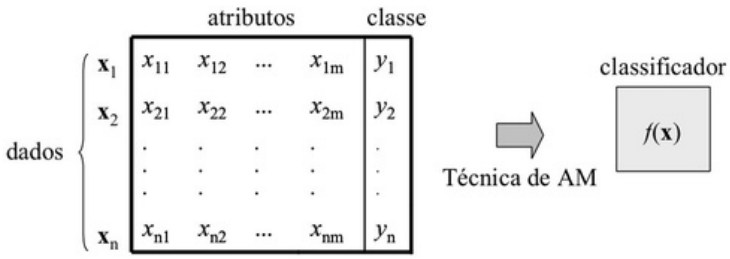


Figura 3 – **Representação esquemática do funcionamento de um algoritmo de aprendizado supervisionado.** Fonte: Extraído de Lorena e Carvalho (2007).

(Figura 3). Neste caso, dado um conjunto rotulados da forma (*xi, yi*), e para cada amostra *xi* com *m* atributos e rótulos/saídas *yi* associados, objetivamos induzir (ou treinar) um modelo - que também pode ser visto como uma função *f* - a partir de uma fatia dos dados (conjunto de treino) e que seja capaz de predizer o rótulo de novas amostras. Quando os desfechos a serem preditos são valores discretos, como diferentes categorias, denominamos como uma tarefa de classificação; quando os desfechos são valores contínuos, denominamos como uma tarefa de regressão.

Modelos baseados em aprendizado supervisionado vêm sendo amplamente utilizados em estudos dentro das grandes áreas de ciências da saúde e da vida para os mais diversos fins (AL-TASHI et al., 2023; MALINDA, 2023; ZHOU et al., 2023; YANG et al., 2023; REEL et al., 2021), devido ao volume massivo de dados produzido com o aprimoramento de técnicas no campo da genética, proteômica e metabolômica. O mesmo pode ser dito para estudos em *C. sativa*, seja para fins industriais ou forenses (AASIM et al., 2022; JAGANNATHAN, 2020; LU et al., 2022; CISANA et al., 2024). Neste trabalho, daremos foco a diferentes algoritmos clássicos de aprendizado supervisionado que serão descritos a seguir, com o objetivo de classificar amostras de *C. sativa* oriundas da rede de tráfico brasileira quanto ao seu local de origem de cultivo a partir de dados de grande dimensionalidade de marcadores do tipo *SNP* e *SSR*, avaliando agora seu potencial para rastreabilidade quando associados à técnicas de *ML*.

2.1 Random Forest

Random Forest (RF) é um método *ensemble* que combina o desfecho de um conjunto de árvores de decisão criadas de forma randômica, gerando uma "floresta", que dá nome ao algoritmo

(BREIMAN, 2001). As árvores de decisão são algoritmos de estrutura hierárquica que se assemelham a fluxogramas de tomadas de decisão, comumente utilizados em modelos de aprendizado supervisionado e sendo a base para algoritmos mais complexos como RF. Os principais elementos de uma árvore de decisão são o nó raiz, os nós internos ou de decisão, os galhos e os nós folha. A ideia principal é representar os dados de forma que, a partir de um nó raiz onde a população inicial de dados irá se dividir, cada nó interno subsequente representa um teste ou uma condição aplicada em um dos atributos ou *features* X, associados as amostras, e cada galho representa um desfecho caso essa condição seja ou não cumprida por este atributo em questão. Estas condições podem ser pontos de corte que separam as diferentes classes em casos de *features* representadas por variáveis contínuas, ou diferentes grupos que fazem o mesmo dentro de variáveis categóricas (*e.g.:* diferentes genótipos em um dado locus). Por fim, os nós folha, também chamados de nós terminais, possuem uma classe Y que será atribuída a amostra testada dependendo de qual caminho de nós foi seguido (Figura W). Para definir qual será a ordem das *features* testadas nos nós internos, utilizamos a métrica de impureza de *Gini*, de forma que quanto menor a impureza da *feature*, melhor sua qualidade para iniciar a árvore e a separação dos dados. Para um nó com K classes, a impureza é dada por:

$$Gini = 1 - \sum_{i=1}^{K} p2i$$

O algoritmo de RF usa um sistema conhecido como *Bootstrapping + AGGregating*, ou *BAGGing*. Em resumo, realiza-se uma amostragem aleatória da população original juntamente a uma seleção aleatória de *features* do conjunto de dados original (*Bootstrapping*) para criar uma série de árvores de decisão únicas, garantindo que cada árvore seja treinada com um subconjunto diferente de amostras e *features*. Depois que todas as árvores são geradas, dados novos ou não vistos pelo modelo são passados por cada árvore e recebem uma classificação. Finalmente, os resultados de classificação de todas as árvores são combinados paralelamente (*Aggregating*) e a decisão final é tomada por meio de um sistema de votação majoritária, onde a classe da amostra será aquela que recebeu mais votos.

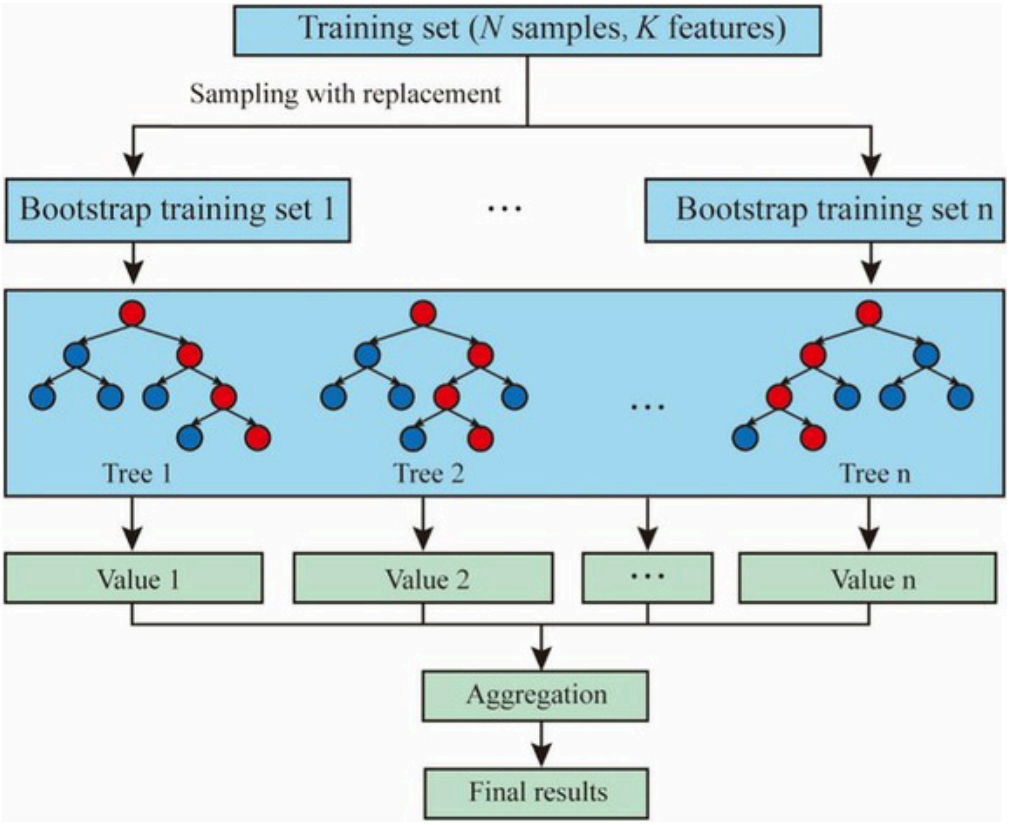


Figura 4 – **Representação gráfica do sistema de *BAGGing* em um algoritmo *Random Forest***.
Fonte: Extraído de Dai et al. (2022).

2.2 Gradient Boosting

Assim como o *RF*, o *Gradient Boosting (GB)* (FRIEDMAN, 1999) é um algoritmo *ensemble* que também se utiliza das estruturas de árvores de decisão, mas diferente do *RF*, baseia-se no conceito de *weak learners* que são somados de forma sequencial e buscam melhorar gradualmente seu desempenho de classificação de forma iterada. Neste contexto, *weak learners* são pequenas árvores de decisão, de um ou dois nós internos, treinadas para reduzir os resíduos da iteração anterior, e aproximando-se cada vez mais da classificação correta. O primeiro passo do algoritmo é o cálculo uma previsão inicial das amostras pertencerem a uma classe *k*, baseado em sua proporção no conjunto de dados, segundo a equação:

$$pk = \frac{\sum i=1 N yi}{N}$$

E a partir desta proporção, é determinada uma previsão inicial (F_k0) em *log-odds* para cada classe,

$$F0 = \log\left(\frac{pk}{1-pk}\right)$$

que após é transformada em um valor de probabilidade inicial ($\hat{p}ki$) utilizando a função *softmax*:

$$\hat{p}ki = \frac{e^{Fk0(xi)}}{\sum_{j=1}^{N} eFj0(xi)}$$

- $\hat{p}ki$ é a probabilidade prevista para classe *k* na amostra *i*;
- $Fk0(xi)$ é a *log-odds* calculada para classe *k* na amostra *i* na *m*-ésima iteração do *GB*;
- *N* é o número total de classes;
- $\sum_{j=1} eFj0(xi)$ somadosexponenciaisdetodasasclasses,queagecomofatordenormalização.

Por fim, essa probabilidade é utilizada para o cálculo do resíduo (rki), que são calculados a partir do gradiente negativo de uma função de perda *log-loss* (*L*) pré-definida, mas pode ser matematicamente simplificadocomoadiferençaentreaclasseverdadeira(yk)eaprevista($\hat{p}k$ i):

$$rk = \frac{)L}{)Fk0(xi)}$$

$$\Downarrow$$

$$rk_i = yk_i - \hat{p}k_i$$

- yki assume o valor de 1 caso a amostra pertença à classe em questão, ou 0 caso não pertença. Este cálculo é realizado para todas as classes;

Aqui entram os *weak learners*, que são treinados para prever estes resíduos, gerando novas funções $hk0(xi)$ que servem como correções da previsão inicial. O modelo *GB* é então atualizado, somando essa correção multiplicada por uma taxa de aprendizado (hiperparâmetro pré-definido) à previsão inicial e gerando uma nova previsão mais próxima da classe real da cada amostra. Estes passos são repetidos até que se atinja um critério de parada, seja ele um número máximo alcançado de árvores ou ausência de melhora na função de perda. Assim, no modelo final, a predição é a soma das funções das árvores:

$$Fk_M(x) = Fk_0(x) + \sum_{m=1}^{M} hkm(xi)$$

A classificação final da amostra será aquela classe com a maior probabilidade $\hat{p}ki$ dentre as classes disponíveis após a finalização das iterações do modelo.

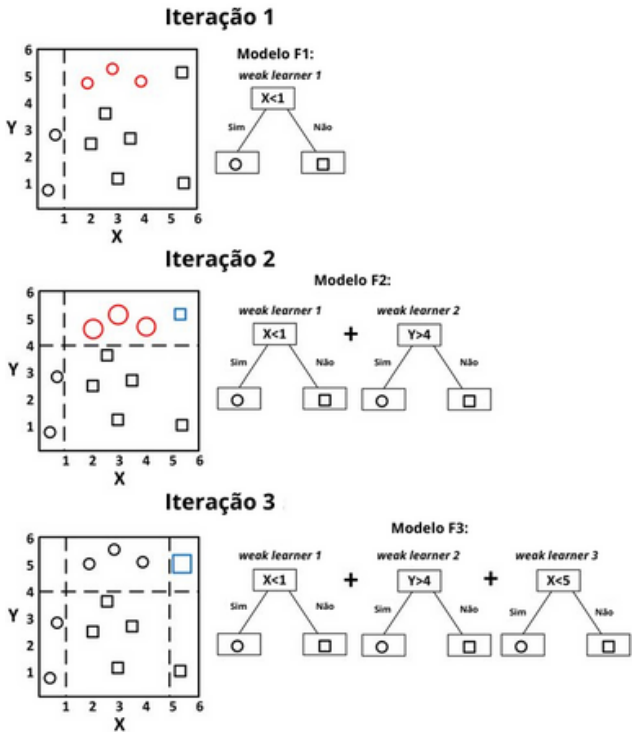


Figura 5 – **Representação gráfica do treinamento sequencial de *weak learners*.** Fonte: Adaptado de Zhang et al. (2018).

2.3 Support Vector Classifier

O *Support Vector Classifier (SVC)* é uma subcategoria do algoritmo denominado *Support Vector Machine (SVM)* (VAPNIK, 1995) voltado a problemas de classificação. O objetivo de um *SVM* é encontrar o hiperplano que ofereça a melhor separabilidade entre as classes em um espaço multidimensional, de forma que em um espaço de *N*-dimensional, o hiperplano será (*N* – 1)-dimensional. O hiperplano é estabelecido de forma que a margem entre duas classes seja maximizada, sendo a margem definida como a distância entre os pontos representantes mais próximos de cada classe, os quais chamamos de vetores de suporte (CERVANTES et al., 2020; CHEN; YIN; TIAN, 2024; LORENA; CARVALHO, 2007).

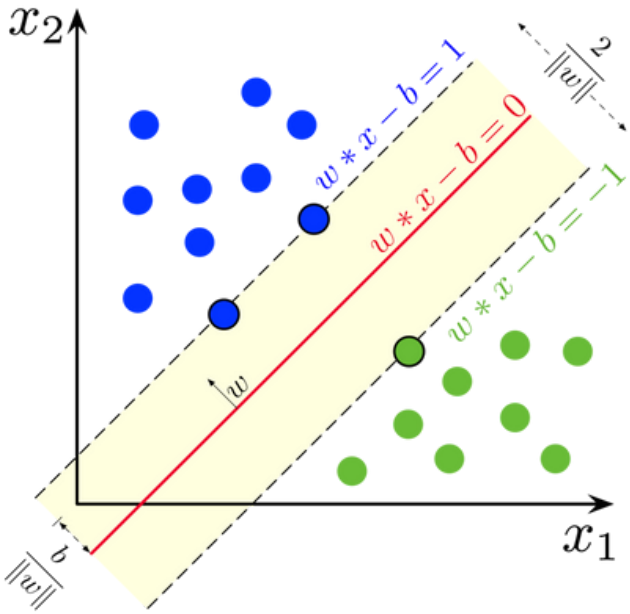


Figura 6 – **Representação gráfica de um hiperplano de margem rígida em um espaço bidimensional.** A linha vermelha equivale ao hiperplano, a margem equivale ao espaço em amarelo e os pontos cruzados pelas retas tracejadas são os vetores de suporte. Fonte: Extraído de Kodur e Naser (2021)

2.3.1 Margem rígida

O melhor hiperplano possível, ou seja, aquele que distingue perfeitamente as duas classes em dados linearmente separáveis, é denominado margem rígida, sendo definido pela equação:

$$F(x)=!1 \cdot x1+!2 \cdot x2...!n \cdot xn+b=0$$

Onde:

- *!* é o vetor de pesos, que define a direção ou inclinação do hiperplano que separa as classes;
- *b* é o *bias* que determina posição do hiperplano;
- *n* é o número total de *features* presentes no conjunto de dados.

Assim, a equação divide o espaço em $! \cdot x + b > 0$ e $! \cdot x + b < 0$ e apresenta uma função de restrição $\forall i = 1$, $yi(! \cdot xi + b) - 1 \geq 0$, ou seja: se a classe de amostra for positiva ($yi = +1$), a saída será de no mínimo 1 e o ponto está no espaço positivo do hiperplano e a pelo menos uma unidade de distância da margem superior, e se a classe da amostra for negativa ($yi = -1$), a saída

será de no mínimo -1 e o ponto estará no espaço negativo do hiperplano e a pelo menos uma unidade de distância da margem inferior. Essa restrição obriga os pontos a estarem fora da margem, o que significa que não são permitidas classificações equivocadas, imaginando que estamos tratando de dados perfeitamente separáveis e sem erros de classificação. Neste contexto de margem rígida, o objetivo do *SVC* é encontrar o hiperplano ótimo que apresente a maior margem possível. Para isso, deseja-se minimizar a norma euclidiana do vetor *!* (||*w*||)-vistoqueamargeméinversamente proporcional.Issosetraduznafunçãoobjetivo:

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

2.3.2 Margens suaves

Não é incomum nos deparamos com diversos casos onde os dados não são perfeitamente separáveis de forma linear, seja pela natureza do problema ou pela presença de *outliers*. Ainda assim, é possível que a implementação de um hiperplano resulte numa boa separação das classes e bom desempenho de classificação. Nestes casos, são necessários ajustes na função de restrição, permitindo que hajam classificações errôneas e que hajam pontos no espaço da margem ou além, tratando-os como exceções controladas. Denominamos este caso de *SVMs* de margem suave. Assim, os dados estarão sujeitos a nova função de restrição:

$$yi(w \cdot xi + b) \geq 1 - i, \quad \forall i = 1, \ldots, n$$

Onde:

- *i* são variáveis de folga que permitem violações da classificação ótima citada anteriormente.
- Se *i* = 0, então a função de restrição *yi*(*!* · *xi* + *b*) – 1 ≥ 0 ainda se aplica;
- Se *i* ≥ 0, o ponto pode estar dentro da margem ou além.

E a nova função objetivo penaliza os desvios das variáveis de folga somando $\sum i$multiplicadopor umhiperparâmetro*C*pré-definidoquecontrolaograudepenalização:

$$\min_{w,b,} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} i$$

Valores maiores de *C* minimizam as violações da função de restrição, tornando-a mais próxima de uma margem rígida, enquanto valores menores toleram mais a presença de *outliers*.

2.3.3 Função *kernel*

Em determinados casos, tanto *SVMs* de margem rígida ou suave não serão capazes de fazer a separação das classes, porque o conjunto de dados não é linearmente separável através de um hiperplano com o número de dimensões disponíveis. Nestes cenários, é possível fazer uso de funções denominadas *kernels*, ou também chamados *kernel tricks*. Para lidar com problemas não lineares, *SVMs* utilizam uma função de mapeamento ( ) dos dados de treinamento de um espaço de entrada X para um espaço de maior dimensão denominado espaço de *features* ((*x*)) (Figura 7), onde assim poderão ser separados por um *SVM* linear pois há uma maior margem de separação.

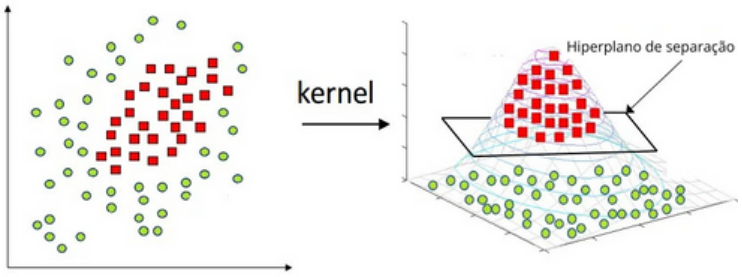


Figura7–**Representação gráfica do aumento de dimensionalidade utilizando *kernel trick*.** Fonte: Extraídoeadaptadode<https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d>.

Onde, por exemplo, é possível levar os dados de um espaço de 2 dimensões ($x1$, $x2$) para umespaçode3dimensõesonde $(x,x)=(x,x,x2+x2$ 2).Oobstáculo,noentanto,resideno fato de que o espaço de *features* pode apresentar um número muito alto de dimensões (até mesmo infinitas) dependendo da transformação aplicada ao espaço de entrada, o que pode tornar o processo computacionalmente custoso. O *kernel trick* entra para resolver este problema, pois, em essência, o seu papel é obter a única informação necessária para a separação linear dos dados, que se trata do produto escalar de dois pontos *xi* e *xj* através de uma função *kernel* (*K*), diretamente do espaço de entrada, evitando que todos estes cálculos sejam computados no espaço de *features* e otimizando a operação. Tem-se, então: $K(x1, x2) = (x1) \cdot (x2)$

Neste trabalho, apenas *SVMs* lineares foram utilizados como modelos de classificação, mas podemos citar como outras funções *kernel* amplamante utilizadas:

- Polinomial: $K(x,y) = (x \cdot y + c)d$, que permite encontrar relações polinomiais entre as variáveis;
- Gaussianaou *RadialBasisFunction(RBF)*:$K(x,y)=\exp-\left(\|x-y\|^2\right)$,transformaosdados emumespaçoinfinitoepodeencontrarrelaçõesmaiscomplexasentrevariáveis.

2.3.4 *SVMs* em problemas multiclasse

SVMs são, por natureza, algoritmos estruturados para classificações binárias. Para problemas multi-classe com *N* classes, duas abordagens podem ser tomadas: (1) método *One-vs-Rest (OvR)*, envolve dividir o problema de classificação em *N* classificadores binários, onde em cada modelo, uma classe específica será diferenciada do restante dos dados e terá uma probabilidade calculada, sendo escolhida aquela com maior probabilidade calculada; (2) método *One-vs-One (OvO)* cria um modelo binário para cada par de classes, gerando $N(N-1)/2$ classificadores, onde o resultado final é decidido por um sistema votação, sendo atribuída a classe mais votada.

2.4 Otimização de hiperparâmetros

Como mencionado em tópicos anteriores, modelos são a combinação de funções matemáticas associadas à parâmetros e hiperparâmetros ajustáveis que auxiliam na melhora de seu desempenho. Ambos são caracterísiticos da natureza de cada modelo, no entanto, os parâmetros são valores internos que são ajustados automaticamente durante o processo de treinamento (*e.g.:* vetor de pesos *!* e *bias* dentro da função determinada por um *SVM*). Os hiperparâmetros, por outro lado, são valores definidos antes do treinamento do modelo e que afetam sua estrutura interna, impactando diretamente seu desempenho final. Podemos citar como exemplo: o número limite de árvores, profundidade e número máximo de nós folha em um *RF*, a taxa de aprendizado em um *GB* e tipo de função *kernel* e parâmetro *C* em um *SVM*. Os hiperparâmetros podem ser definidos de forma manual ou utilizando

estratégias de otimização, que irão encontrar a combinação de hiperparâmetros que oferecerá o melhor desempenho para o modelo (FEURER; HUTTER, 2019). São as principais delas:

2.4.1 *Grid Search*

É a estratégia mais simples de otimização, porém, a mais custosa computacionalmente. Dado um conjunto ou faixa de valores atribuídos a um ou mais hiperparâmetros, o algoritmo de *Grid Search* irá testar todas as possíveis combinações de valores e retornar a combinação cuja métrica avaliativa de escolha apresentou o melhor desempenho.

2.4.2 *Random Search*

Semelhante à estratégia anterior, porém realiza um número especificado de combinações aleatórias de valores atribuídos a um ou mais hiperparâmetros, e não todas as combinações possíveis. É uma alternativa para tarefas com um número elevado de amostras ou muitos hiperparâmetros a serem ajustados simultâneamente.

2.4.3 Otimização Bayesiana

A otimização bayesiana é uma abordagem mais eficiente para a busca de hiperparâmetros, utilizando um modelo probabilístico, geralmente um processo Gaussiano, para estimar uma função de perda e selecionar os próximos pontos a serem avaliados. Esse processo é guiado por uma função de aquisição, que equilibra a exploração de regiões desconhecidas e a exploração dos melhores candidatos já identificados. Em outras palavras, o algoritmo identifica as regiões do espaço de busca de valores pré-estipulados para cada hiperparâmetro e identifica onde estão os maiores valores de performance, dando preferência para valores próximos desta região e aumentando as chances de obter uma melhor performance. A cada iteração, os novos hiperparâmetros testados refinam o modelo probabilístico, tornando a busca melhor direcionada. Dessa forma, reduz significativamente o número de avaliações, sendo o mais indicado para tarefas com um grandes *datasets* ou muitos hiper parâmetros.

2.5 Validação cruzada

A validação cruzada é a estratégia utilizada para avaliar o desempenho e a capacidade de generalização dos modelos preditivos, onde busca-se dividir o conjunto de dados em um grupo de treino e outro grupo de teste do modelo. Caso o modelo seja treinado e testado em um mesmo conjunto, podemos observar um fenômeno conhecido como *overfitting*, onde os dados de treinamento são memorizados e não ocorre o aprendizado de um padrão generalizável. Existem diferentes estratégias de validação cruzada que permitem diferentes graus de aproveitamento de um *dataset*, sendo eles:

2.5.1 *Hold-out validation*

Consiste na simples divisão dos dados em dois subconjuntos, um de treino e outro voltado para teste do modelo. Uma proporção comumente utilizada é de 70% usado para treino e 30% utilizado para testar a generalização do modelo. A desvantagem está na "perda" de uma parcela dos dados que irá ser destinada ao grupo teste. Além disso, é necessário o cuidado de implementação da distribuição estratificada das amostras de diferentes classes, garantindo que todas as classes presentes terão a mesma parcela de amostras presentes nos dois grupos.

2.5.2 *K-fold*

Nesta estratégia, os dados são divididos em *K* partes ou *folds*. Assim, *K* – 1 *folds* serão utilizadas para treinamento, enquanto o *fold* restante será utilizado para testar o modelo. O processo irá se repetir *K* vezes até que todas as partes sejam utilizadas tanto para treino quanto para teste. Assim, o desempenho médio do modelo é calculado e irá oferecer uma noção mais abrangente de sua capacidade preditiva.

2.5.3 *Leave-One-Out*

Semelhante à estratégia anterior, com a diferença de que no *Leave-One-Out (LOO)*, cada amostra servirá invididualmente como ponto de teste enquanto as demais estarão reservadas para treino. Assim, o *LOO* trata-se de um *K-fold* aonde *K* é igual ao número de amostras disponíveis.

Fornece a avaliação mais detalhada possível da capacidade do modelo, sendo bastante custosa em grandes *datasets*, mas muito vantajosa em *datasets* pequenos.

2.6 Seleção de Atributos

Seleção de atributos (*FS*, do inglês *Feature Selection*) refere-se a diferentes metodologias utilizadas para redução de dimensionalidade em grandes conjuntos de dados (BUTVINIK, 2021). De forma sucinta, o objetivo do processo de *FS* é maximizar a relevância e reduzir a redundância dos dados, identificando e removendo atributos (ou *features*) ruidosas, redundantes ou pouco informativas. Trata-se de uma etapa fundamental do tratamento de dados em larga escala, principalmente em cenários onde encontramos a problemática do "*large p, small n*", ou seja, o número de observações (*n*) é milhares ou até centenas de milhares vezes menor do que o número de *features* (*p*) relacionadas. Tal cenário vêm se tornado comum especialmente dentro de estudos ômicos e multi-ômicos (BALLARD et al., 2024; HUYNH; NGUYEN; DO, 2020; MEI; WANG, 2016) devido aos avanços na área da biotecnologia e biologia molecular, e dentro do contexto do *ML*, isso reflete no que chamamos de "Maldição da dimensionalidade" - isto é, à medida em que a dimensionalidade (representada pelo número de *features*) aumenta, o número de observações necessárias para o bom desempenho de qualquer algoritmo de *ML* aumenta exponencialmente (ALTMAN; KRZYWINSKI, 2018; BELLMAN, 1961). Desta forma, a aplicação de *FS* auxilia não apenas na melhora de desempenho dos modelos gerados, mas também na redução de fenômenos como *overfitting* e na redução do custo computacional. Como métodos comumente utilizados para *FS* dentro do aprendizado supervisionado, podemos citar:

2.6.1 Métodos *Filter*

Utilizam de testes estatísticos (*e.g.:* coeficientes de correlação, filtros de variância, teste de Chi-quadrado, entre outros) para ranquear e selecionar as *features* mais relevantes, sendo independentes de algoritmos de aprendizado. São especialmente úteis em conjuntos massivos de dados, devido a seu baixo custo computacional.

2.6.2 Métodos *Wrapper*

Algoritmos *Wrapper* tratam o processo de *FS* como uma tarefa de busca, onde diferentes combinações de *features* são selecionadas, ranqueadas e comparadas com outras combinações, através de modelos preditivos internos. Por conta disso, são os métodos computacionalmente mais custosos, mas que tendem a oferecer melhores resultados. Alguns exemplos são:

- *Forward/Backward Feature Selection*: as *features* são adicionadas (*Forward*) ou removidas (*Bakward*) uma a uma, de forma iterativa, de acordo com seu acréscimo na performance do modelo interno utilizado para avaliá-las.

- *Exhaustive Feature Selection*: faz uma busca exaustiva dentre todas as combinações possíveis de subconjuntos de *features*, retornando a que apresentou melhor desempenho no modelo inter no.

- *Recursive Feature Elimination*: faz a remoção de *features* de forma recursiva até que a quantidade desejada seja obtida, isto é: o modelo interno é treinado com *x features* e remove aquela que apresentou menor importância dentro de sua contrução, sendo novamente treinado com este novo subconjunto de $x - 1$ *features*, repetindo o processo até que a quantidade pré-determinada seja atingida.

2.6.3 Métodos *Embedded*

Diferente dos anteriores, métodos *Embedded* incorporam o *FS* no processo de treinamento de seu modelo interno, realizando-os de forma simultânea, semelhante ao funcionamento das estruturas de árvore de decisão presentes nos algoritmos de *RF* e *GB*. Isso lhes possibilita melhor ajustar as *features* à complexidade do modelo interno utilizado, diferente dos métodos *Filter*, ao mesmo tempo em que os torna computacionalmente menos custosos se comparados à métodos *Wrapper*, oferecendo assim "o melhor dos dois mundos".

2.7 Visão geral

Na segunda sessão deste capítulo, buscou-se compreender a definição de *ML*, com foco na grande área do aprendizado supervisionado, assim como explorar com maior profundidade o funcionamento, a organização estrutural e a lógica-matemática que baseia os modelos adotados no presente trabalho, sendo eles: *Random Forest*, *Gradient Boosting* e *Support Vector Classifier*. Finalmente, tratamos dos principais conceitos e estratégias utilizadas dentro do *ML* para buscar a melhor construção possível e o bom desempenho de um modelo preditivo, como a otimização dos hiperparâmetros, a validação cruzada e, por último, o conceito de seleção de atributos, essencial para redução da dimensionalidade dos dados e, consequentemente, sua redundância, com destaque para métodos *Filter*, *Wrapper* e *Embedded*, todos utilizados em etapas posteriores do desenvolvimento do presente trabalho.

CAPÍTULO 2: Objetivos

1. Objetivo Geral

O objetivo do presente trabalho é a construção de modelos preditivos capazes determinar a origem de cultivo de amostras de *C. sativa* oriundas da rede de tráfico brasileira, a partir de dados genotípicos de marcadores *SNP* e *SSR*, utilizando de técnicas de redução de dimensionalidade e algoritmos clássicos de aprendizado de máquina supervisionado.

2. Objetivos específicos

2.1 Identificar e extrair variantes *SNP* e *SSR* a partir de dados alinhados de sequenciamento completo de amostras de *C. sativa*

- Extração de *SNPs* através de um *pipeline* adaptado de chamada de variantes;
- Extração de *SSRs* através do *software SATIN*;
- Codificação e transformação dos dados genotípicos para conformação *feature matrix* e posterior utilização em técnicas de *ML*.

2.2 Reduzir a dimensionalidade dos dados a partir de diferentes estratégias de seleção de atributos;

- Obtenção de subconjuntos distintos das variantes de maior relevância selecionados através de cada estratégia de seleção utilizada;

2.3 Avaliar o desempenho de diferentes combinações de modelos preditivos de classificação induzidos através de algoritmos clássicos de *ML* a partir dos subconjuntos de variantes obtidos.

2.4 Investigar a correlação biológica entre marcadores *SNP* e *SSR* selecionados pelos modelos com melhor desempenho e sua acurácia.

Capítulo 3:

Apresentação do artigo científico

O artigo entitulado "Identification of biogeographically informative microssatelite markers for Brazilian Cannabis sativa samples: a Machine Learning approach for forensic origin prediction" foi resultado de parte do trabalho desenvolvido durante o período do curso de Mestrado, e foi formatado seguindo as normas do periódico *International Journal of Legal Medicine*, para o qual foi submetido para publicação.

Identification of biogeographically informative microssatelite markers for Brazilian Cannabis sativa samples: a Machine Learning approach for forensic origin prediction

Cá´ssioAugustoRodriguesBettim1,2*,
Lucas de Oliveira Pereira Ribeiro2,
Oscar Victor Cardenas Alegŕı´a5, Eduardo Filipe Avila Silva2,
Franciele Maboni Siqueira4, Clarice Sampaio Alho2,5,
Má´rcioDorn1,2,3,4

[1*]Laboratory of Structural Bioinformatics and Computational Biology, Federal University of Rio Grande do Sul, Av. Bento Gonç¸alves 9500, Porto Alegre, 91509-900, RS, Brazil.
[2]National Institute of Forensic Science and Technology, Porto Alegre, RS, Brazil.
[3]Institute of Informatics, Federal University of Rio Grande do Sul, Av. Bento Gonç¸alves 9500, Porto Alegre, 91509-900, RS, Brazil.
[4]Center for Biotechnology, Federal University of Rio Grande do Sul, Av. Bento Gonç¸alves 9500, Porto Alegre, 91509-900, RS, Brazil.
[5]Institute of Biological Sciences, Federal University of Pará´, Av. GeneraÍ´ssimoDeodoro92,Belé´m,66050-160,PA,Brazil.
[5]Pathology Post Graduation Program, Health Sciences Federal University of Porto Alegre, R. Sarmento Leite 245, Porto Alegre, 90050-170, RS, Brazil.

Contributing authors: mdorn@inf.ufrgs.br;

Abstract

In Brazil, Cannabis sativa is the most commonly trafficked illicit drug, with supply chains originating from both domestic and international sources, often

Capítulo 4:
Predição de origem a partir de variantes do tipo *SNP* utilizando aprendizado de máquina supervisionado

1. Materiais e métodos

1.1 Amostras

Um total de 38 amostras de *C. sativa* foram disponibilizadas pela Polícia Federal Brasileira (*BFP*, do inglês *Brazilian Federal Police*) de quatro grupos diferentes e divididas de acordo com seu local de origem de cultivo: Paraguai (*PG*), Colômbia (*CO*), região do Polígono da Maconha (*MP*, do inglês *Marijuana Polygon*) e amostras de cepas comerciais estrangeiras de origem não especificada, consideradas como Grupo Estrangeiro (*FG*, do inglês *Foreign Group*). As amostras *PG* (n= 8) e *CO* (n=4) foram obtidas de sementes presentes em tijolos de cannabis - folhas secas, flor e caule prensados juntos - apreendidos por meio de triagem postal entre 2019 e 2022. As amostras *MP* (n= 12) foram obtidas de material vegetal foliar apreendido in loco em operações da BFP entre os anos de 2015 a 2019. As amostras *FG* (n=14) também foram obtidas de material vegetal foliar, cultivado a partir de sementes de diferentes cepas comerciais apreendidas nos serviços postais em 2017 e 2019 e plantadas para testes de confirmação pela *BFP*.

1.2 Extração, quantificação e sequenciamento

O isolamento de DNA foi realizado usando o Dneasy mericon Food Kit (Qiagen, Valencia, CA) para amostras de pequena escala e o *DNeasy Plant Mini Kit* (Qiagen, Valencia, CA) para folhas

e sementes, respectivamente, seguindo o protocolo do fabricante, com exceção de: 20 mg de material foliar inicial usado para o Dneasy mericon Food Kit e uma semente por amostra usada para o DNeasy Plant Mini Kit. As concentrações genéticas do material foram quantificadas pelo *Qubit dsDNA HS Assay Kit* (Invitrogen, Carlsbad, Califórnia, EUA). O sequenciamento das 38 amostras foi realizado na plataforma *Illumina NovaSeq* 6000 com leitura paired-end de 250 pb. A demultiplexação de todas as bibliotecas para cada faixa de sequenciamento foi realizada através do software *Illumina bcl2fastq v2.20.*

1.3 Controle de qualidade

O controle de qualidade dos 38 arquivos de sequenciamento (*SRA*) foi realizado utilizando as ferramentas FastQC v0.11.9 (ANDREWS, 2010) e, quando necessário, Trimmomatic v0.39 (BOLGER; LOHSE; USADEL, 2014) para remoção de possíveis contaminantes e ou adaptadores presentes, com a seguinte ordem de passos de trimagem e parâmetros para Paired-End Data: ILLUMINACLIP, utilizando um arquivo fasta contendo sequências específicas illumina, SLIDINGWINDOW:4:15 e MINLEN:75.

1.4 Alinhamento

Os arquivos fastq foram mapeados contra o genoma de referência cs10 (GCA_900626175.2 *GenBank acession number*) (GRASSA et al., 2018) utilizando o algoritmo alinhamento BWA-MEM2 v2.2 (MD et al., 2019). Os arquivos BAM resultantes foram classificados por coordenada e reads duplicadas marcadas utilizando a ferramenta *Picard v2.9.4* (Broad Institute, 2018) através das linhas de comando *SortSam* e *MarkDuplicates*, respectivamente. Métricas de alinhamento foram obtidas utilizando *samtools v1.17* (LI et al., 2009) flagstat.

1.5 Chamada de variantes e filtragem

O processo de chamada de variantes para os arquivos BAM de cada amostra foi realizado utilizando a função *HaplotypeCaller* da ferramenta GATK v4.2.0 (MCKENNA et al., 2010), gerando

arquivos no formato *Variant Call Format* (*VCF*) para cada amostra. Em seguida, as variantes do tipo *SNP* foram extraídas dos *VCFs* resultantes por meio da função *SelectVariants* do GATK v4.2.0, e filtrados manualmente utilizando a função *VariantFiltration* com os seguintes critérios: *QD* < 2, *FS* > 60, *MQ* < 40, *MQRankSum* < −12.5 e *ReadPosRankSum* < −8.

1.6 Recalibração de bases e *Bootstraping*

Os arquivos *VCFs* contendo as variantes filtradas foram utilizados para o passo de recalibração de bases (*BQSR*, do inglês *Base Quality Score Recalibration*), recomendado pelo manual de boas práticas do GATK. Como não há disponível um set de variantes conhecidas para nosso organismo de estudo, realizamos um processo conhecido como *Bootstraping*. O processo consiste em performar o passo de chamada de variantes sem realização do *BQSR* e filtrá-las manualmente para obter um conjunto de variantes de alta confiabilidade, o que já foi feito no passo anterior. Este set de variantes servirá como conjunto de variantes conhecidas para a realização do *BQSR* nos arquivos BAM de cada amostra. Primeiramente, a função *BaseRecalibrator* GATK v4.2.0 cria uma tabela de recalibração para *BQSR* baseada em diversas covariáveis. No segundo passo, esta tabela é utilizada para aplicação da *BQSR* em nossos dados através da função *ApplyBQSR* GATK v4.2.0, gerando arquivos BAM recalibrados, dos quais uma nova chamada de variantes é realizada através da função *HaplotypeCaller* GATK v4.2.0, gerando novos VCFs definitivos de cada amostra.

1.7 Criação do banco de dados genômico e filtragem final

A função *GenomicDBImport* GATK v4.2.0 é utilizada para criação de um banco de dados genômicos unindo todas as amostras, agora separadas por cromossomo. Por último, a função *GenotypeGVCF* GATK v4.2.0 é utilizada para genotipagem conjunta de todas as amostras a partir do banco de dados de cada cromossomo, gerando arquivos *VCFs* genotipados. Estes *VCFs* contendo todas as variantes detectadas passam novamente para a separação dos SNPs e filtragem dos mesmos através da funções *SelectVariant* e *VariantFiltration* seguindo os mesmos passos e utilizando os mesmos critérios mencionados anteriormente, resultando nos *VCFs* finais.

1.8 Filtragens adicionais

Para fins de redução de dimensionalidade dos arquivos e prosseguimento do *pipeline*, uma filtragem adicional utilizando a ferramenta *Plink* v2.0 (CHANG et al., 2014) foi realizada, com os filtros –maf 0.05 e –geno 0.1. Para isso, a nomenclatura dos cromossomos dos arquivos VCF foi adaptada para o formato válido para a ferramenta e os sítios multialélicos foram divididos em múltiplas linhas com a ferramenta bcftools v1.17 (DANECEK et al., 2021) através das funções *annotate* e *norm*, respectivamente.

1.9 Imputação de dados faltantes e *Feature Selection*

Sabendo que a presença de dados faltantes podem introduzir vieses no processo de FS para redução de dimensionalidade dos dados (SEIJO-PARDO et al., 2019), a ferramenta *Beagle* v5.4 (BROWNING; ZHOU; BROWNING, 2018) foi utilizada para imputação de dados faltantes nos arquivos VCF. Após o processo de imputação, a etapa de seleção de atributos foi feita através do software *Variant Spark* (BAYAT et al., 2020), que baseia-se no modelo de *Random Forest* (*RF*) para identificar quais atributos - ou variantes - contribuíram e em qual proporção para um determinado desfecho a ser previsto, sendo uma ferramenta altamente apropriada para dados genômicos de alta dimensionalidade. Para a etapa de *FS*, o arquivo *VCF* foi convertido no formato *TSV* (*Table Separated Values*) e adaptado para conformação *Feature matrix*. Além disso, os genótipos foram codificados representando a contagem do alelo alternativo (e.g.: AA = 0, AC = 1, CC = 2), considerando o efeito aditivo em cada variante. A partir deste arquivo, foram extraídas as 30 variantes que mais contribuíram para a tarefa de predição da origem geográfica das amostras, baseado no critério de *Gini Importance*.

1.10 Tarefa de classificação quanto ao local de origem

Para determinar a origem das amostras de *C. sativa*, partindo do painel de 30 variantes SNP extraídas na etapa anterior, três modelos de classificadores diferentes foram explorados e avaliados: *Support Vector Classifier* (*SVC*), *Random Forest (RF) Classifier* e *Gradient Boosting (GB)*

Classifier. O *kernel* linear foi definido anteriormente para o modelo SVC. Pesos de classe balanceados foram definidos para os modelos *SVC* e *RF* para lidar com o desequilíbrio de classes. O ajuste de hiperparâmetros para cada um dos modelos citados foi realizado anteriormente usando otimização bayesiana com 100 iterações, validação cruzada de 3 vezes e *F1-score* ponderado como função de pontuação. Os modelos otimizados foram avaliados usando *Leave-One-Out Cross-Validation (LOO-CV)* e, para determinar o melhor número e combinação de *SNPs* para classificação das amostras dentro de cada modelo, o método de eliminação recursiva de atributos (*RFE*, do inglês *Recursive Feature Elimination*) foi aplicado, onde o número de variantes utilizado era aumentado em ordem crescente (1 a 30 variantes). Para cada variante adicionada desta forma, métricas ponderadas de Acurácia (*Acc*), *F1-Score (F1)*, Sensibilidade (*Rec*) e Precisão (*Prec*) foram computadas, sendo feitas 30 réplicas em cada ponto. O número ideal de variantes de cada modelo foi definido quando o pico de desempenho era atingido com o menor número de variantes necessárias para tal.

1.11 Análises adicionais de *clustering*

Como complemento, e para cada modelo de classificação, análises de *clustering* foram realizadas em *Python* 3.10.12 usando o número ótimo e a combinação de SNPs com o qual o maior *F1-score* foi alcançado. A Análise de Coordenadas Principais (*PCoA*) foi conduzida utilizando a métrica *cityblock* (*Manhattan distance*) usando as bibliotecas *scipy v1.11.2* e *scikit-bio v0.5.8*. O *clustering* hierárquico, usando o método de grupo de pares não-ponderados com média aritmética (*UPGMA*) e métrica *cityblock* (*Manhattan distance*), foi conduzido usando a biblioteca *scipy v1.11.2*.

1.12 Análises adicionais de parâmetros forenses e índices de genética de populações

Os parâmetros forenses, utilizados para avaliação dos marcadores para fins investigativos, e índices de genética de populações foram calculados utilizando a ferramenta online *STRAF v2.2.2* (GOUY; ZIEGER, 2017) com a adaptação dos arquivos para dados do tipo SNP, sendo eles: Heterozigosidade esperada (*Hexp*), Heterozigosidade observada (*Hobs*), Poder de discrminação (*PD*), Poder de exclusão (*PE*), Probabilidade de coincidência (*MP*), Conteúdo de informação sobre polimorfismo (*PIC*), Frequências alélicas (*AF*) e Desequilíbrio de ligação (*LD*) entre as variantes.

2. Resultados adicionais e Discussão

Depois de realizados todos os processos de filtragem, obtivemos o arquivo VCF final com 1667500variantesidentificadas.Análisesdadistribuiçãodetiposdesubstituiçãoerazão*Ts/Tv*acima de 0,5, variando entre 1.8 e 1.9 (Figura 8) são indicativos de SNPs de maior qualidade identificados através do *Variant Calling*, embora esses limites possam variar entre espécies. Alguns fatores metodológicos podem ter causado a diminuição dessas estatísticas: o uso da técnica *bootstrapping* para executar o *BQSR* e a necessidade de usar o método de filtragem manual para selecionar os *SNPs* de maior qualidade. Finalmente, após a imputação de dados faltantes, a etapa de *FS* realizada pelo software *Variant Spark* identificou os 30 *SNPs* mais relevantes (Material suplementar - Tabela 1) durante o processo de redução de dimensionalidade, as quais foram subsequentemente extraídas do *VCF* original e utilizadas para a tarefa de classificação nos modelos *SVC*, *RF* e *GB*.

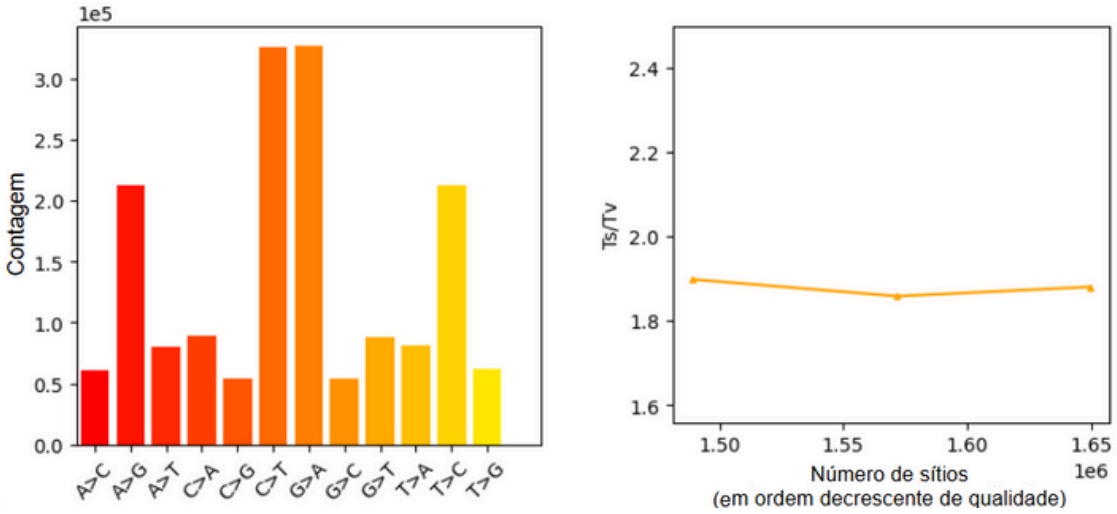


Figura 8 – **Tipos de substituições (à esquerda) e razão de transições e transversões (à direita) observadas no conjunto final de variantes extraídas.**

Quanto à tarefa de predição de origem e classificação das amostras de *C. Sativa*, o destacou-se o modelo SVC de kernel linear, que apresentou excelente desempenho e alcançando o *F1-Score* máximo de 1.0 (Figura 9) utilizando apenas 5 SNPs, sendo eles - por ordem decrescente de relevância: CHR 1 POS 44446322 G>T, CHR 8 POS 31428091 C>G, CHR 4 POS 84487478 C>T, CHR 8 POS 51750844 T>A e CHR 8 POS 51751977 G>T. Esses resultados indicam que tais marcadores são extremamente discriminativos e fornecem excelente separabilidade linear para distinção entre

as amostras, mesmo utilizando-se de um modelo de classificação mais simples, de baixo custo computacional e facilmente interpretável. Os modelos baseados em GB e RF também mostraram ótima capacidade preditiva, apresentando valores de métricas avaliativas máximas muito semelhantes, utilizando-se de listas com mais variantes em seu pico de desempenho - F1: 0.99; Acc: 0.99; Prec: 0.99; Rec: 0.99 utilizando 7 SNPs e F1: 1.00; Acc: 1.00; Prec: 1;00; Rec: 1.00 utilizando 14 SNPs, respectivamente (Figuras 10 e 11). Curiosamente, as variantes CHR 1 POS 44446322 G>T, CHR 4 POS 84487478 C>T e CHR 8 POS 51751977 G>T foram compartilhadas entre as listas de variantes utilizadas pelos três modelos em seus picos de performance, porém em posições de relevância diferentes dentro de cada algoritmo. Isso indica que os três SNPs possuem uma importância geral para classificação das amostras e sugere uma possível relação entre as variantes e o desfecho, ou seja, o local de origem dos diferentes grupos de *C. sativa*. As diferentes posições de relevância das variantes estabelecidas em cada tarefa de classificação pode ser explicada não somente pelas diferenças de mecanismo de avaliação de features de cada algoritmo, mas também por diferenças na interação destas com outros SNPs, visto que a importância de cada feature não é determinada de forma isolada.

CHR 1 POS 44446322 G>T, o marcador de maior relevância, está presente exclusivamente em amostras do grupo MP e está localizado em uma região não codificante, downstream do gene LOC115706163, que codifica um fator de transcrição *DPB* (*DNA Polymerase Binding*), e upstream do pseudogene LOC115704083. As proteínas DPB, juntamente com os fatores de transcrição E2F, formam um heterodímero responsável pela modulação de transcrição de diversos genes responsáveis pelo controle positivo e negativo do ciclo celular mediado por fosforilação via CDK, divisão celular e endoreplicação em *Arabidopsis thaliana* (*A. thaliana*) (KOSUGI; OHASHI, 2002; POZO et al., 2006; POZO et al., 2007). O segundo SNP de maior relevância, CHR 8 POS 31428091 C>G, presente em amostras dos grupos MP e CO, está localizado no gene LOC115700822, que codifica uma proteína ainda não caracterizada. O terceiro SNP de maior importância, CHR 4 POS 84487478 C>T, também está presente em amostras do grupo MP e CO e localizado em uma região não codificante, downstream dos genes LOC115714397, que codifica as isoformas X1 e X2 da *U-Box domain-containing protein 5*, e LOC115714483, que codifica uma proteína ainda não caracterizada. Proteínas *U-Box* são E3 ligases com domínio altamente conservado e responsáveis por modificações pós-traducionais na via ubiquitina-proteassoma, mecanismo essencial de degradação de proteínas

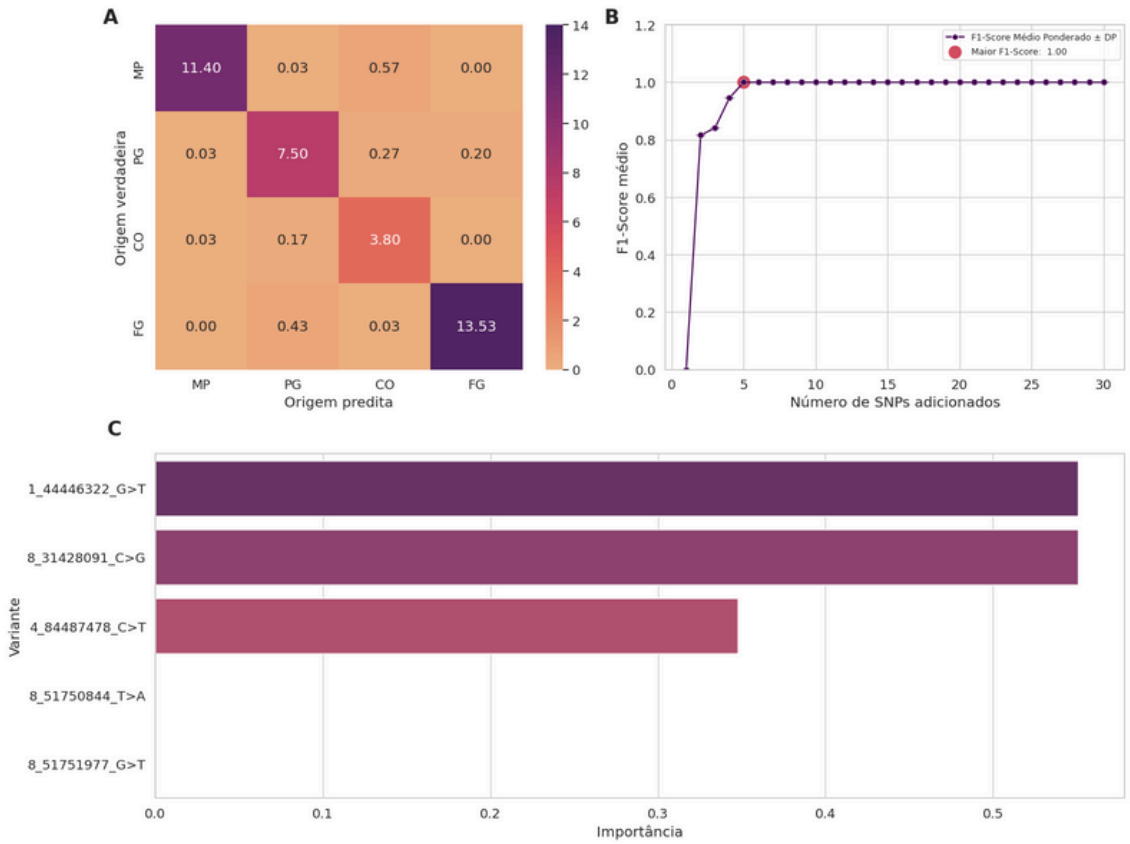


Figura 9 – **Avaliação dos Modelos de classificação baseados em SVC utilizando *kernel* linear.** (A) Mostra a matriz de confusão média em todas as réplicas, ilustrando o desempenho da classificação nos quatro grupos de origem: MP, PG, CO e FG. (B) Curva de aprendizado de *F1-score* ponderado (com desvios padrão) como uma função do número de marcadores SNPs adicionados. O ponto vermelho marca o *F1-score* mais alto observado. (C) Importâncias dos 5 SNPs usados no modelo com melhor desempenho.

e relacionado com diversas respostas à fatores externos e estresses abióticos em plantas, como a seca (ANDERSEN et al., 2004; AZEVEDO; SANTOS-ROSA; SHIRASU, 2001; KIM; KIM; SEO, 2023; TRENNER et al., 2022). Por fim, os quarto e quinto SNPs de maior relevância para o modelo, CHR 8 POS 51750844 T>A e POS 51751977 G>T, também estão localizados em uma região não codificante, downstream do gene LOC115698661 e upstream do gene LOC115701657, que codificam uma histona do tipo 4 (*H4*) e *small nucleolar RNA R71*, respectivamente. Estes são os dois únicos SNPs presentes em amostras do grupo PG tanto em estado homozigoto quanto heterozigoto, mas também estão presentes em amostras dos grupos MP e CO, o que é um indicativo de que podem ter sido utilizadas para identificação de plantas oriundas do Paraguai dentro do modelo de classificação e pode explicar sua menor relevância dentro da tarefa de separação. Estudos prévios já indicaram que

mutações na cauda e domínio globular de H4 aceleram o ciclo de floração em *A. thaliana*, indicando um papel importante da histona na regulação do tempo de florescimento (CORCORAN et al., 2022; SUNDAR et al., 2024).

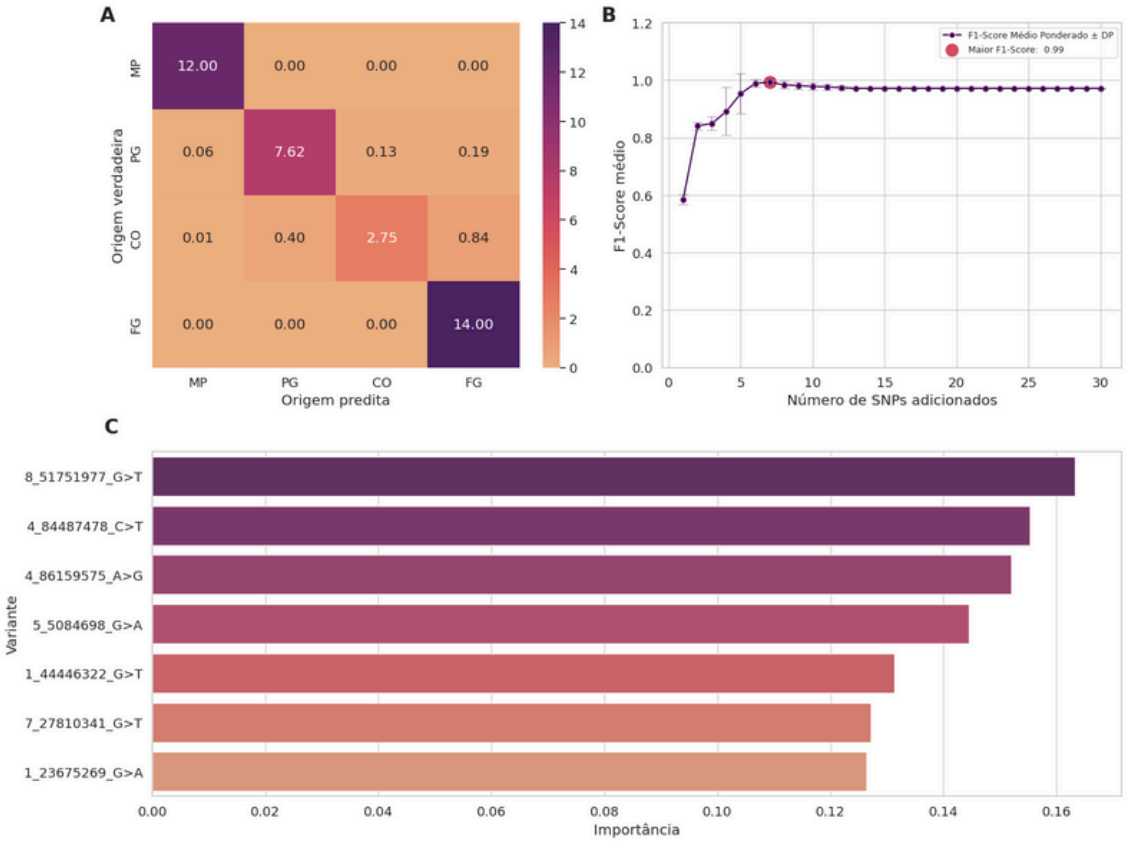


Figura 10 – **Avaliação dos Modelos de classificação baseados em GB.** (A) Mostra a matriz de confusão média em todas as réplicas, ilustrando o desempenho da classificação nos quatro grupos de origem: MP, PG, CO e FG. (B) Curva de aprendizado de *F1-score* ponderado (com desvios padrão) como uma função do número de marcadores SNPs adicionados. O ponto vermelho marca o *F1-score* mais alto observado. (C) Importâncias dos 7 SNPs usados no modelo com melhor desempenho.

Apesar da ausência de estudos prévios citando os SNPs identificados, os resultados do modelo de predição sugerem um possível papel discriminatório desses marcadores para a separação das populações analisadas. O SNP CHR 1 POS 44446322 G>T, presente exclusivamente na classe MP, pode ser indicativo de diferenças populacionais específicas, possivelmente relacionadas ao histórico de cultivo ou adaptação ambiental local. Embora esteja localizado em uma região não codificante, sua proximidade com o gene LOC115706163 (fator de transcrição DPB) levanta a hipótese de um potencial impacto regulatório indireto na expressão gênica, especialmente considerando o papel

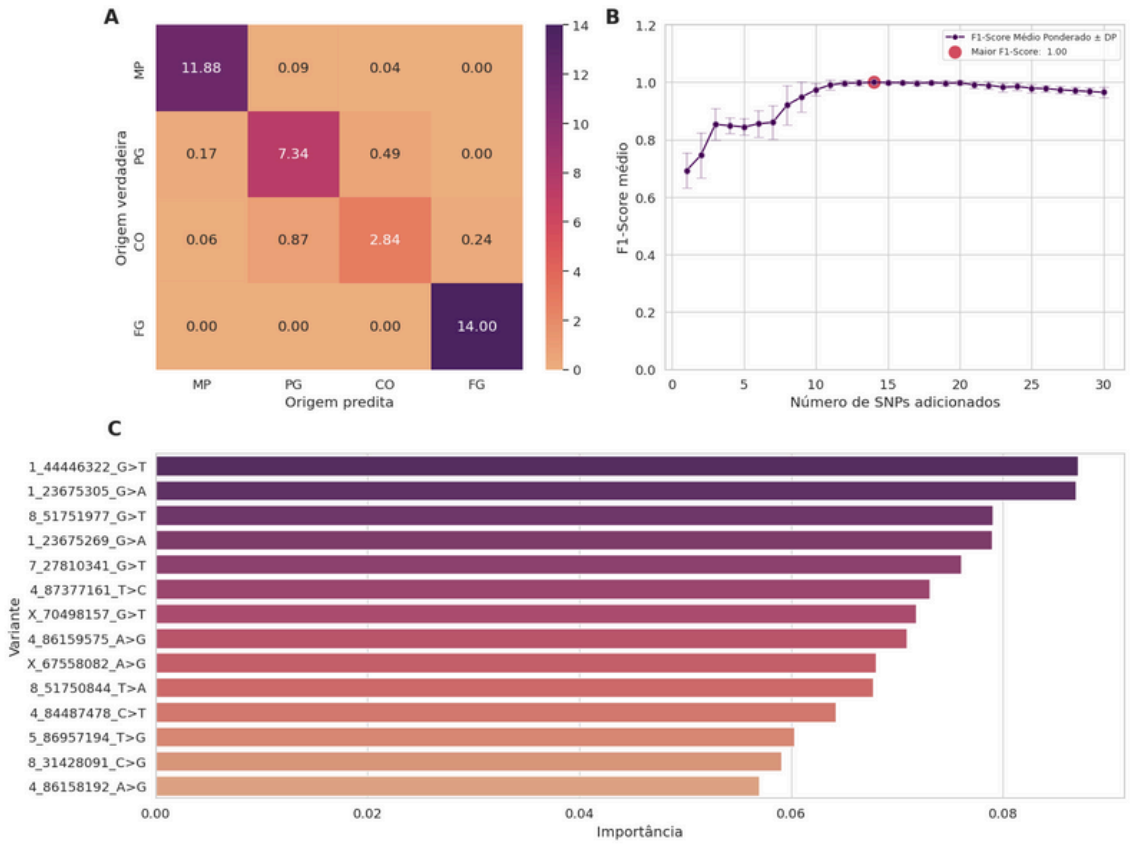


Figura 11 – **Avaliação dos Modelos de classificação baseados em RF.** (A) Mostra a matriz de confusão média em todas as réplicas, ilustrando o desempenho da classificação nos quatro grupos de origem: MP, PG, CO e FG. (B) Curva de aprendizado de *F1-score* ponderado (com desvios padrão) como uma função do número de marcadores SNPs adicionados. O ponto vermelho marca o *F1-score* mais alto observado. (C) Importâncias dos 14 SNPs usados no modelo com melhor desempenho.

deste gene na modulação do ciclo celular em *A. thaliana*. Os SNPs CHR 8 POS 51750844 e POS 51751977, localizados em regiões não codificantes associadas ao gene da histona H4, sugerem um possível envolvimento no controle do tempo de floração, como observado em outras plantas modelo, ainda que sem validação funcional em *C. sativa*. Além disso, a presença desses SNPs em múltiplos grupos populacionais pode refletir variabilidade genética compartilhada ou fluxo gênico entre as populações estudadas. Embora os dados estejam limitados a análises preditivas, os resultados apresentados fornecem uma base para futuros estudos de validação e aplicação prática, como o uso dos SNPs como marcadores moleculares para identificação populacional ou seleção genética.

As matrizes de desequilíbrio de ligação (LD) entre os SNPs selecionados (Figura 12) foram analisadas com o objetivo de identificar variantes que possam ter sido herdadas em conjunto, tornando-

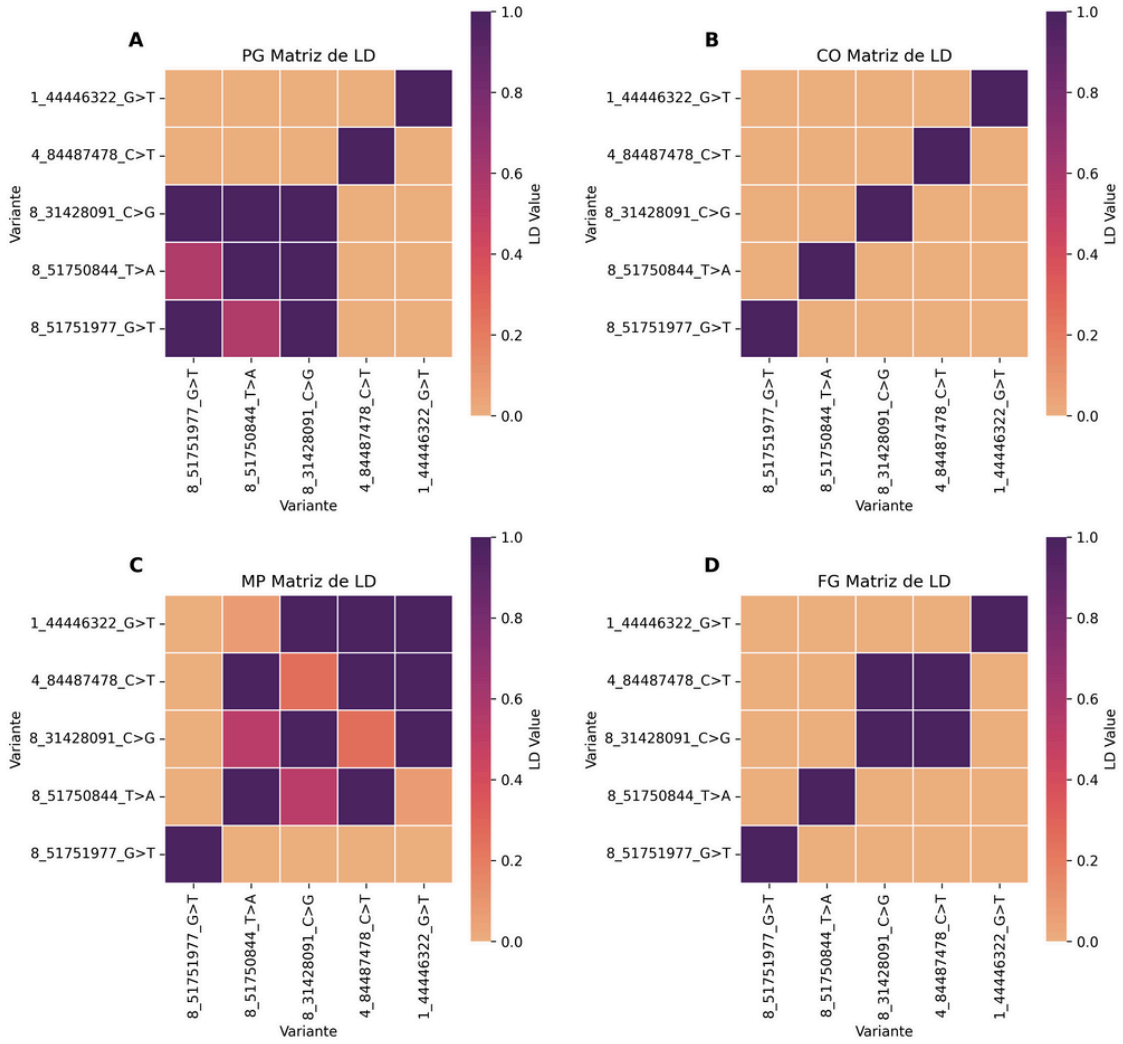


Figura 12 – **Matrizes de desequilíbrio de ligação (LD) entre os 5 SNPs selecionados pelo modelo *SVC* através dos diferentes grupos de origem.** (A) Amostras do Paraguai (PG). (B) Amostras da Colombia (CO). (C) Amostras do Polígono da Maconha (MP). (D) Amostras do grupo estrangeiro (FG).

se possivelmente redundantes dentro da classificação. Os cinco SNPs selecionados revelam diferentes padrões de associação genética entre os grupos analisados. No grupo PG, destacam-se fortes associações (r^2=1.0) entre os SNPs localizados no cromossomo 8 (CHR 8 POS 31429091 C>G, CHR 8 POS 51758044 T>A e CHR 8 POS 51751977 G>T), sugerindo uma redução recombinacional devido a proximidade física dos marcadores e, além disso, que essas variantes estão ligadas e podem não contribuir de forma independente para a discriminação entre as amostras dentro do grupo, mas que podem ter sido úteis para classificação de outras amostras. Já no grupo MP, observa-se um padrão mais amplo de LD elevado (r^2=1.0), envolvendo SNPs localizados em diferentes cromossomos (CHR

SNP	Hexp	Hobs	PIC	PM	PD	PE
CHR1POS44446322G>T	0.380	0.013	0.304	0.519	0.480	0.013
CHR4POS84487478C>T	0.336	0.421	0.277	0.512	0.487	0.127
CHR8POS31428091C>G	0.478	0.236	0.360	0.375	0.624	0.040
CHR8POS51750844T>A	0.489	0.447	0.366	0.369	0.630	0.145
CHR8POS51751977G>T	0.497	0.131	0.370	0.403	0.596	0.013

Tabela 1 – **Parâmetros forenses calculados para os 5 SNPs selecionados.** São eles Heterozigosidade esperada (Hexp), Heterozigosidade observada (Hobs), Conteúdo de Informação Polimórfica (PIC), Probabilidade de correspondência aleatória (PM), Poder de Discriminação (PD) e Poder de Exclusão (PE).

SNP	REF	ALT
CHR1POS44446322G>T	0.750	0.250
CHR4POS84487478C>T	0.789	0211
CHR8POS31428091C>G	0.618	0.382
CHR8POS51750844T>A	0.592	0.408
CHR8POS51751977G>T	0.434	0.566

Tabela 2 – **Frequências alélicas dos 5 SNPs selecionados.** Referentes aos alelos de referência (REF) e alternativos (ALT).

1, CHR 4 e CHR 8), o que pode indicar a presença de haplótipos conservados ou eventos de seleção que mantiveram essas variantes em associação. Adicionalmente, valores intermediários de r^2 no grupo MP (como 0.5 e 0.25) sugerem uma conexão moderada entre certos loci, podendo refletir um grau maior de recombinação. No grupo FG, a forte associação entre CHR 4 POS 84487478 C>T e CHR 8 POS 31428091 C>G (r^2=1.0) reforça a hipótese de herança conjunta dessas variantes, enquanto no grupo CO a ausência de valores elevados de LD sugere uma maior diversidade genética ou maior ocorrência de recombinação. Esses resultados indicam que algumas variantes selecionadas pelo modelo, especialmente no cromossomo 8, podem ser altamente correlacionadas e não agregar informações independentes para a classificação dentro de determinados grupos, mas ao mesmo tempo possuem relevância para separação das amostras no demais. No entanto, é necessário cuidado ao considerar valores elevados de LD considerando tamanho amostral reduzido dos grupos, podendo tratar-se de um artefato estatístico.

As frequências alélicas (Tabela 2) dos cinco SNPs selecionados também foram utilizadas para o cálculo dos parâmetros forenses (Tabela 1). Para avaliar a informatividade dos marcadores, foram levados em consideração os valores de PD, PE, PIC, Hexp e Hobs. Os SNPs considerados mais informativos foram CHR 8 POS 51750844 T>A e CHR 8 POS 51751977 G>T, cujo desempenho

	CO	PG	MP	FG
FG	0.6566	0.7122	0.7966	-
MP	0.3546	0.4943	-	
PG	0.2041	-		
CO	-			

Tabela 3 – ***Fixation index (FST)* entre os quatro grupos de origem.** Distância genética calculada a partir dos genótipos dos 5 SNPs selecionados.

ainda assim pode ser interpretado como moderado para fins de identificação. Curiosamente, tais marcadores foram os que apresentaram menor relevância dentro da tarefa de classificação das amostras no modelo SVC. É importante salientar que tais parâmetros são projetados para avaliar a eficácia de um marcador genético na identificação individual. Em contrapartida, a predição de origem por ML foca na atribuição de uma amostra a uma classe ou grupo de origem em específico. Um marcador pode ser informativo para diferenciação individual de amostras de *C. sativa*, mas se os alelos desse *SNP* forem distribuídos de forma semelhante entre as diferentes classes, ele terá baixa relevância para a classificação. Por outro lado, um *SNP* fixado ou com baixa variabilidade (por exemplo, presente em apenas uma classe) pode ser altamente relevante para a classificação, mesmo com valores modestos nestes parâmetros forenses.

Como validações complementares, métodos de agrupamento adicionais de *PCoA* (Figura 13) e clusterização hierárquica (Figura 14) foram efetuados, também a partir dos genótipos das amostras referentes aos 5 SNPs selecionados pelo modelo de classificação em seu pico de desempenho. A distância genética entre os quatro grupos de origem também foi determinada através do cálculo de índice de fixação (*FST*), apresentado na tabela 3. No PCoA, as PC1 e PC2 explicam, respectivamente, 81,12% e 9,10% da variância total entre as amostras, sendo capaz de segregar os quatro grupos no espaço bidimensional, mesmo. É possível observar certos padrões consistentes ao comparar os resultados das três análises. O grupo FG apresentou os maiores valores de *FST* em relação aos demais (0,7966 em relação a MP e 0,7122 em relação a PG), indicando um alto grau de diferenciação genética, o que foi refletido na clara separação observada tanto no dendrograma quanto no *PCoA*. Por outro lado, CO e PG mostraram menor distância genética (*FST* = 0.2041), sugerindo maior similaridade genética, o que também foi evidenciado pela proximidade desses grupos no espaço bidimensional do PCoA e pelo agrupamento mais próximo no dendrograma. A posição intermediária do grupo MP em relação a PG (*FST*=0.4943) e CO (*FST*=0.3546) reforça a transição genética entre

esses grupos, sugerindo possíveis relações evolutivas ou históricos genéticos compartilhados.

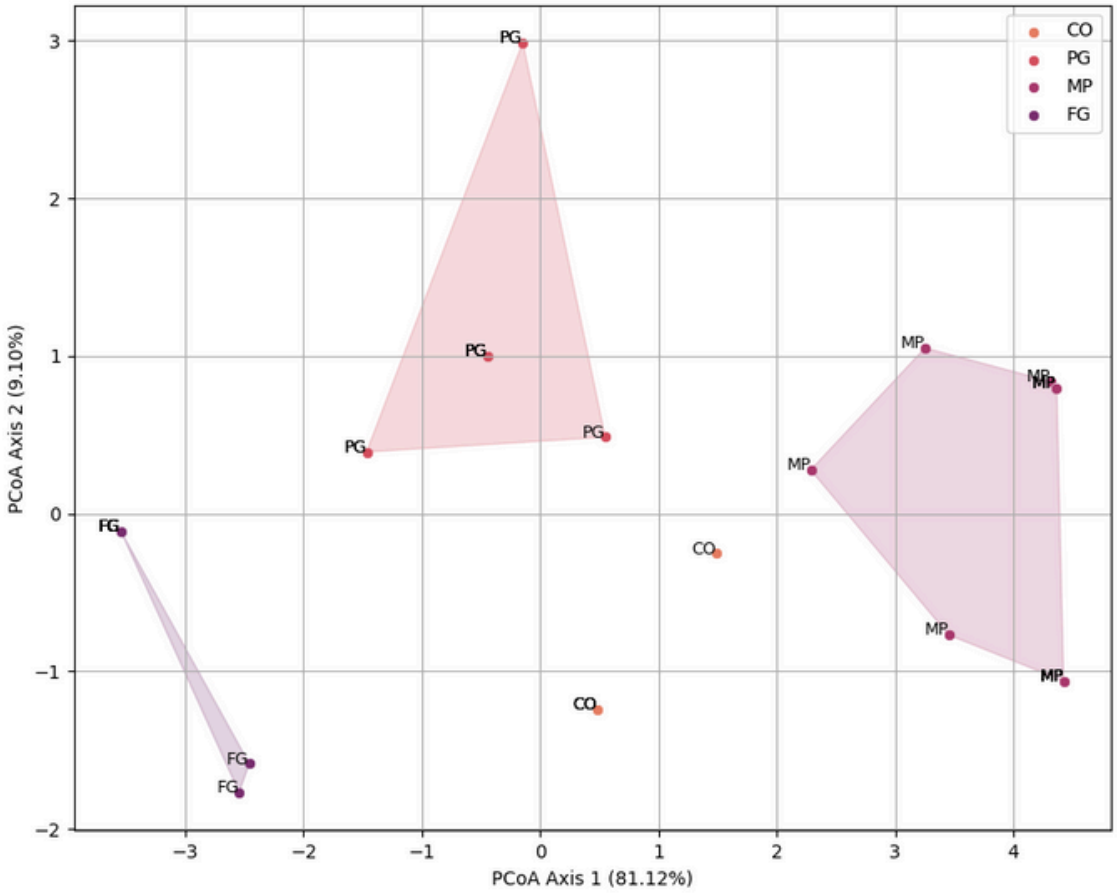


Figura 13 – ***Principal Coordinate Analysis (PCoA)*** **baseada na distância de Manhattan entre as amostras.** Cálculo de distância realizado a partir do genótipo dos 5 SNPs elencados pelo melhor modelo de classificação *SVC* para todo o *dataset* ordenado nas dimensões PC1 e PC2, que explicam 81,12% e 9,10%, respectivamente.

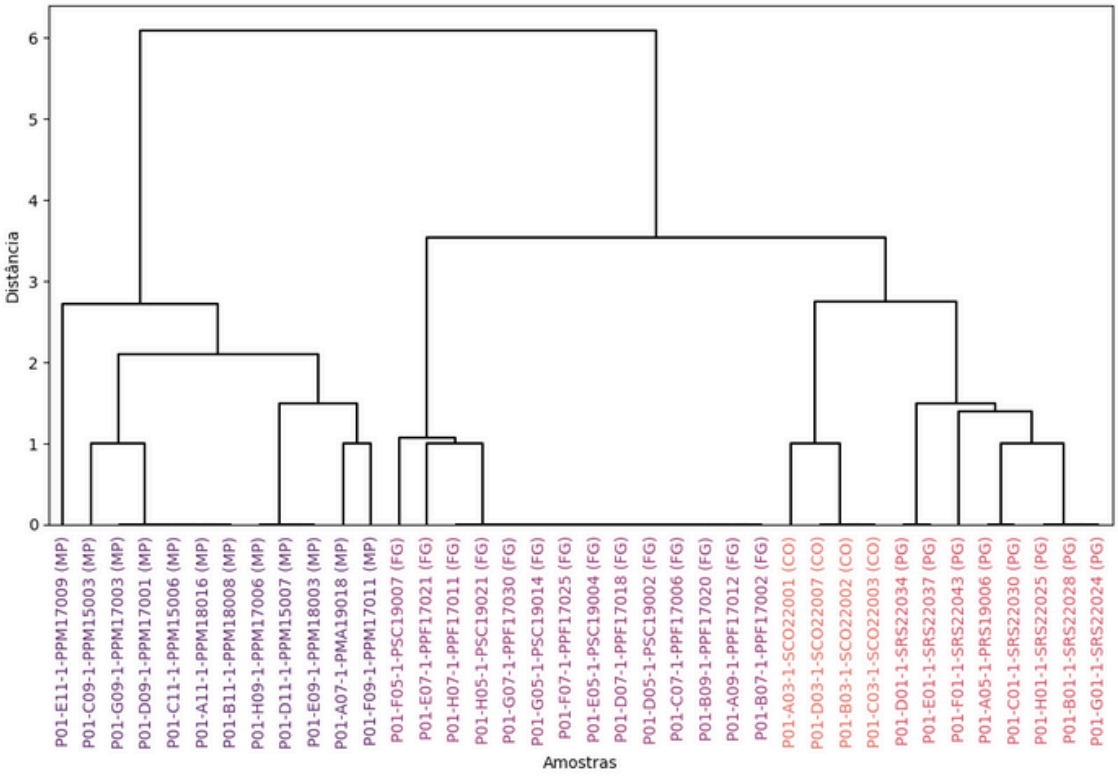


Figura 14 – ***Hierarchical Clustering*** **baseado no método *UPGMA* e distância de Manhattan.** O dendrograma ilustra as relações genéticas entre as amostras a partir do genótipo dos 5 SNPs selecionados pelo melhor modelo de classificação *SVC*. As amostras estão identificadas a partir de cores referentes ao seu grupo de origem, sendo MP em roxo escuro, PG em vermelho, CO em laranja e FG em roxo

Capítulo 5: Conclusão geral

Em suma, concluímos que tanto a combinação da informação genotípica de diferentes marcadores *SNPs* quanto *SSRs*, quando associados à diferentes estratégias clássicas de *ML*, são capazes de atuar como variáveis informativas para a indução de modelos de classificação relativos a origem biogeográfica de amostras de *C. sativa*, reforçando seu potencial de rastreabilidade para fins investigativos já observado em literatura. Além disso, foi possível observar que, tratando-se de modelos baseados em *SVC* - associados à técnicas distintas de redução de dimensionalidade -, alcançamos a separabilidade perfeita dos diferentes grupos de origem biogeográfica utilizando subconjuntos consideravelmente reduzidos de ambos os marcadores quando comparados à painéis utilizados em estudos prévios combinados à técnicas de clusterização para diferenciação tanto de amostras brasileiras quanto estrangeiras. Os modelos de classificação baseados nos algoritmos de *RF* e *GB* também apresentaram excelentes resultados quanto a predição de origem das amostras, no entanto, demonstraram necessitar de subconjuntos maiores de marcadores *SNP* e *SSR* para alcançar o mesmo desempenho. Ao mesmo tempo, classificadores *GB* mostraram instabilidade de performance em suas métricas avaliativas tanto na comparação de seus diferentes picos de desempenho associados a metodologias distintas de *FS* - variando de 80% a até 100% de acerto -, quanto durante a construção das curvas de aprendizado utilizadas para definir o número ótimo de *features* a serem utilizadas.

Apesar do resultados promissores, uma limitação significativa deste trabalho reside no número reduzido amostras disponíveis, ainda que proveninentes das principais rotas de tráfico do país, o que pode impactar a capacidade de generalização dos classificadores construídos. Estudos envolvendo *ML* apresentam uma quantidade bastante variável de número amostral disponível para treinamento, comumente situado entre centenas e milhares de pontos de observação a depender do domínio de pesquisa. Amostras limitadas podem levar a um sobreajuste dos modelos aos dados de treinamento, o *overfitting*, reduzindo sua confiabilidade e capacidade generalização quando aplicado a novos dados, mesmo com a implementação de técnicas como validação cruzada. Além disso, estratégias de *FS*, embora úteis para identificar os marcadores mais informativos, podem ter sua eficácia comprometida em conjuntos de dados pequenos, aumentando o risco de viés na escolha das *features* consideradas como mais importantes. Assim, futuras investigações com um maior número

de amostras serão essenciais para validar e aprimorar a confiabilidade dos modelos na predição.

Por último, mas não menos importante, podemos citar a limitação relativa à lacunas e genes ainda não caracterizados presentes no genoma de referência da *C. sativa*, o que torna mais difícil estabelecer uma possível relação entre os marcadores levantados por cada modelo preditivo e sua função biológica. Também, embora versões recentes como a cs10 e ASM2916894v1 - com 31.170 e 35.194 genes catalogados, respectivamente - representem um avanço significativo no mapeamento de seu genoma, a presença de regiões altamente repetitivas e mal caracterizadas, especialmente em áreas ricas em microssatélites e variantes estruturais complexas, impacta diretamente estudos baseados em extração de variantes resultando em possíveis falsos positivos e negativos. Não obstante, tais sequências de referência representam apenas determinados cultivares e não refletem totalmente a diversidade genética da espécie, o que pode introduzir viés populacional na predição da origem biogeográfica.

Ainda assim, os resultados deste estudo destacam a relevância da abordagem pioneira de integração de técnicas de aprendizado de máquina supervisionado com marcadores moleculares forenses consagrados, reforçando seu potencial para aprimorar a rastreabilidade e a identificação de amostras de *C. sativa* em investigações criminais e no combate ao tráfico e crime organizado, além de abrir novas perspectivas para aplicações forenses baseadas em inteligência artificial.

Referências

AASIM, M. et al. Machine learning (ml) algorithms and artificial neural network for optimizing in vitro germination and growth indices of industrial hemp (cannabis sativa l.). *Industrial Crops and Products*, Elsevier BV, v. 181, p. 114801, jul. 2022. ISSN 0926-6690. Disponível em: <http://dx.doi.org/10.1016/j.indcrop.2022.114801>. Citadonapágina25.

ABDURAKHMONOV, I. Y. *Microsatellite Markers*. Rijeka: IntechOpen, 2016. ISBN 978-953-51-2798-7. Disponível em: <https://doi.org/10.5772/62560>. Citado na página 22.

ACOSTA, A.; ALMIRALL, J. Differentiation between hemp-type and marijuana-type cannabis using the fast blue bb colorimetric test. *Forensic Chemistry*, Elsevier BV, v. 26, p. 100376, dez. 2021. ISSN 2468-1709. Disponível em: <http://dx.doi.org/10.1016/j.forc.2021.100376>. Citado na página 16.

Agência Nacional de Vigilância Sanitária (ANVISA). *Resolução da Diretoria Colegiada - RDC nº 17, de 06 de maio de 2015*. 2015. Acesso em: 6 mar. 2025. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2015/rdc0017_06_05_2015.pdf>. Citadona página 20.

Agência Nacional de Vigilância Sanitária (ANVISA). *Resolução da Diretoria Colegiada - RDC nº 327, de 9 de dezembro de 2019*. 2019. Acesso em: 6 mar. 2025. Disponível em: <https://www.in.gov.br/en/web/dou/-/resolucao- da- diretoria- colegiada- rdc- n- 327- de- 9- de- dezembro- de- 2019- 232669072> . Citado na página 20.

AL-TASHI, Q. et al. Machine learning models for the identification of prognostic and predictive cancer biomarkers: A systematic review. *International Journal of Molecular Sciences*, MDPI AG, v. 24, n. 9, p. 7781, abr. 2023. ISSN 1422-0067. Disponível em: <http://dx.doi.org/10.3390/ijms24097781>. Citadonapágina25.

ALTMAN, N.; KRZYWINSKI, M. The curse(s) of dimensionality. *Nature Methods*, Springer Science and Business Media LLC, v. 15, n. 6, p. 399–400, maio 2018. ISSN 1548-7105. Disponível em: <http://dx.doi.org/10.1038/s41592-018-0019-x>. Citado na página 36.

ANDERSEN, P. et al. Structure and biochemical function of a prototypical arabidopsis u-box domain. *Journal of Biological Chemistry*, Elsevier BV, v. 279, n. 38, p. 40053–40061, set. 2004. ISSN 0021-9258. Disponível em: <http://dx.doi.org/10.1074/jbc.M405057200>. Citado na página 86.

ANDERSON, L. C. Leaf variation among cannabis species from a controlled garden. *Botanical Museum Leaflets, Harvard University*, Harvard University Herbaria, v. 28, n. 1, p. 61–69, 1980. ISSN 00068098. Disponível em: <http://www.jstor.org/stable/41762825>. Citado 2 vezes nas páginas 15 and 16.

ANDREWS, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Citado na página 80.

ARAUJO, M.; ALMEIDA, M. B.; ARAUJO, L. L. N. The cannabinoids mechanism of action: an overview. *Brazilian Journal Of Pain*, v. 6, 2023. ISSN 2595-0118. Disponível em: <http://dx.doi.org/10.5935/2595-0118.20230028-en>. Citadonapágina18.

ASSIS, B. de et al. The cross-border spillover effects of cannabis regulation in south america. *Journal of Quantitative Criminology*, Springer Science and Business Media LLC, mar. 2025. ISSN 1573-7799. Disponível em: <http://dx.doi.org/10.1007/s10940-025-09603-z>. Citado na página 21.

ASSUNÇAO, M. R. Popular culture and regional society in nineteenth-century maranhao, brazil. *Bulletin of Latin American Research*, Elsevier BV, v. 14, n. 3, p. 265–286, set. 1995. ISSN 0261-3050. Disponível em: <http://dx.doi.org/10.1016/0261-3050(94)00039-J>. Citado na página 14.

AZEVEDO, C.; SANTOS-ROSA, M. J.; SHIRASU, K. The u-box protein family in plants. *Trends in Plant Science*, Elsevier BV, v. 6, n. 8, p. 354–358, ago. 2001. ISSN 1360-1385. Disponível em: <http://dx.doi.org/10.1016/s1360-1385(01)01960-4>. Citadonapágina86.

BALLARD, J. et al. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Mining*, Springer Science and Business Media LLC, v. 17, n. 1, out. 2024. ISSN 1756-0381. Disponível em: <http://dx.doi.org/10.1186/s13040-024-00391-z>. Citado na página 36.

BARBARIć, L.; BEZBRADICA, S. C. A forensic application of genetic markers for distinction between drug-type and fiber-type cannabis sativa l. *Forensic Science International*, Elsevier BV, v. 353, p. 111853, dez. 2023. ISSN 0379-0738. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2023.111853>. Citadonapágina22.

BASTOS, F. I. P. M. et al. *III Levantamento Nacional sobre o Uso de Drogas pela População Brasileira*. [S.l.], 2017. Disponível em: <https://www.icict.fiocruz.br/>. Acesso em: 6 mar. 2025. Citado na página 19.

BAYAT, A. et al. Variantspark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *GigaScience*, Oxford University Press (OUP), v. 9, n. 8, ago. 2020. ISSN 2047-217X. Disponível em: <http://dx.doi.org/10.1093/gigascience/giaa077>. Citado na página 82.

BELLMAN, R. E. *Adaptive Control Processes: A Guided Tour*. [S.l.]: Princeton University Press, 1961. Citado na página 36.

BILBAO, A.; SPANAGEL, R. Medical cannabinoids: a pharmacology-based systematic review and meta-analysis for all relevant medical indications. *BMC Medicine*, Springer Science and Business Media LLC, v. 20, n. 1, ago. 2022. ISSN 1741-7015. Disponível em: <http://dx.doi.org/10.1186/s12916-022-02459-1>. Citado na página 19.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, Oxford University Press (OUP), v. 30, n. 15, p. 2114–2120, abr. 2014. ISSN 1367-4803. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btu170>. Citado na página 80.

BREIMAN, L. Random forests. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1010933404324>. Citadonapágina26.

Broad Institute. *Picard: A Set of Command Line Tools for Manipulating High-Throughput Sequencing (HTS) Data and Formats Such as SAM/BAM/CRAM and VCF*. 2018. <http://broadinstitute.github.io/picard/>. Citadonapágina80.

BROWNING, B. L.; ZHOU, Y.; BROWNING, S. R. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, Elsevier BV, v. 103, n. 3, p. 338–348, set. 2018. ISSN 0002-9297. Disponível em: <http://dx.doi.org/10.1016/j.ajhg.2018.07.015>. Citado na página 82.

BUTVINIK, D. *Feature Selection — Extended Overview*. 2021. Acesso em: 25 mar. 2025. Disponível em: <https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c>. Citadonapágina36.

CARLINI, E. A. A história da maconha no brasil. *J. Bras. Psiquiatr.*, FapUNIFESP (SciELO), v. 55, n. 4, p. 314–317, 2006. Citado na página 14.

CASTILLO-CARNIGLIA, A. et al. Trends in marijuana use in two latin american countries: an age, period and cohort study. *Addiction*, Wiley, v. 115, n. 11, p. 2089–2097, abr. 2020. ISSN 1360-0443. Disponível em: <http://dx.doi.org/10.1111/add.15058>. Citado na página 19.

CERVANTES, J. et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, Elsevier BV, v. 408, p. 189–215, set. 2020. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2019.10.118>. Citado na página 29.

CHANG, C. C. et al. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, v. 4, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:15089556>. Citado na página 82.

CHAYASIRISOBHON, S. Mechanisms of action and pharmacokinetics of cannabis. *The Permanente Journal*, The Permanente Federation, v. 25, n. 1, p. 1–3, mar. 2021. ISSN 1552-5775. Disponível em: <http://dx.doi.org/10.7812/TPP/19.200>. Citado na página 18.

CHEN, X.; YIN, Z.; TIAN, H. Support vector machines principles and actually example. *Procedia Computer Science*, Elsevier BV, v. 243, p. 2–11, 2024. ISSN 1877-0509. Disponível em:

<http://dx.doi.org/10.1016/j.procs.2024.09.002>. Citado na página 29.

CISANA, S. et al. An initial exploration of machine learning for establishing associations between genetic markers and thc levels in cannabis sativa samples. *Forensic Science International: Genetics*, Elsevier BV, v. 73, p. 103123, nov. 2024. ISSN 1872-4973. Disponível em: <http://dx.doi.org/10.1016/j.fsigen.2024.103123>. Citadonapágina25.

CLARKE, R. C.; MERLIN, M. D. *Cannabis: Evolution and Ethnobotany*. 1. ed. University of California Press, 2013. ISBN 9780520270480. Disponível em: <http://www.jstor.org/stable/10.1525/j.ctt3fh2f8>. Citadonapágina13.

CORCORAN, E. T. et al. Systematic histone h4 replacement in arabidopsis thaliana reveals a role for h4r17 in regulating flowering time. *The Plant Cell*, Oxford University Press (OUP), v. 34, n. 10, p. 3611–3631, jul. 2022. ISSN 1532-298X. Disponível em: <http://dx.doi.org/10.1093/plcell/koac211>. Citadonapágina87.

CROCQ, M.-A. History of cannabis and the endocannabinoid system. *Dialogues in Clinical Neuroscience*, Informa UK Limited, v. 22, n. 3, p. 223–228, set. 2020. ISSN 1958-5969. Disponível em: <http://dx.doi.org/10.31887/DCNS.2020.22.3/mcrocq>. Citado na página 13.

CULL, A.; JOLY, D. L. Development and validation of a minimal snp genotyping panel for the differentiation of cannabis sativa cultivars. *BMC Genomics*, Springer Science and Business Media LLC, v. 26, n. 1, jan. 2025. ISSN 1471-2164. Disponível em: <http://dx.doi.org/10.1186/s12864-025-11263-z>. Citado na página 23.

DAI, Y. et al. A hybrid metaheuristic approach using random forest and particle swarm optimization to study and evaluate backbreak in open-pit blasting. *Neural Computing and Applications*, Springer Science and Business Media LLC, v. 34, n. 8, p. 6273–6288, jan. 2022. ISSN 1433-3058. Disponível em: <http://dx.doi.org/10.1007/s00521-021-06776-z>. Citado na página 27.

DANECEK, P. et al. Twelve years of samtools and bcftools. *GigaScience*, Oxford University Press (OUP), v. 10, n. 2, jan. 2021. ISSN 2047-217X. Disponível em: <http://dx.doi.org/10.1093/gigascience/giab008>. Citado na página 82.

DEGENHARDT, F.; STEHLE, F.; KAYSER, O. The biosynthesis of cannabinoids. In: *Handbook of Cannabis and Related Pathologies*. Elsevier, 2017. p. 13–23. ISBN 9780128075630. Disponível em: <http://dx.doi.org/10.1016/b978-0-12-800756-3.00002-8>. Citado na página 17.

DEVINSKY, O. et al. Cannabidiol: Pharmacology and potential therapeutic role in epilepsy and other neuropsychiatric disorders. *Epilepsia*, Wiley, v. 55, n. 6, p. 791–802, maio 2014. ISSN 1528-1167. Disponível em: <http://dx.doi.org/10.1111/epi.12631>. Citado na página 18.

FARRELLY, K. N. et al. The impact of recreational cannabis legalization on cannabis use and associated outcomes: A systematic review. *Substance Abuse: Research and Treatment*, SAGE Publications, v. 17, jan. 2023. ISSN 1178-2218. Disponível em:

<http://dx.doi.org/10.1177/11782218231172054>. Citadonapágina19.

FETT, M. S. et al. 13-loci str multiplex system for brazilian seized samples of marijuana: individualization and origin differentiation. *International Journal of Legal Medicine*, Springer Science and Business Media LLC, v. 133, n. 2, p. 373–384, out. 2018. ISSN 1437-1596. Disponível em: <http://dx.doi.org/10.1007/s00414-018-1940-3>. Citado 2 vezes nas páginas 20 and 22.

FETT, M. S. et al. Geographic origin determination of brazilian cannabis sativa l. (marihuana) by multi-element concentration. *Forensic Science International*, Elsevier BV, v. 315, p. 110459, out. 2020. ISSN 0379-0738. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2020.110459>. Citado na página 20.

FETT, M. S. et al. The use of vegetal tissue multi-element content as an indicator of soil or substrate type employed to cultivate cannabis sativa l. (marijuana). *Forensic Chemistry*, Elsevier BV, v. 23, p. 100319, maio 2021. ISSN 2468-1709. Disponível em: <http://dx.doi.org/10.1016/j.forc.2021.100319>. Citado na página 21.

FEURER, M.; HUTTER, F. *Hyperparameter Optimization*. Cham: Springer International Publishing, 2019. 3–33 p. ISBN 978-3-030-05318-5. Disponível em: <https://doi.org/10.1007/978-3-030-05318-5_1>. Citadonapágina34. FILHO, C. D. et al. *Introdução à Genética Forense: Ciência contr*

o Crime. Campinas, SP: Millenium Editora Ltda., 2020. Citado 2 vezes nas páginas 22 and 23.

FRAGA, P. Rationality of the repression of illicit cannabis crops in brazil: public drug policies and strategies of repression. *Sociology International Journal*, MedCrave Group, LLC, v. 3, n. 2, abr. 2019. ISSN 2576-4470. Disponível em: <http://dx.doi.org/10.15406/sij.2019.03.00180>. Citado na página 20.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 1999. Citado na página 27.

G1. *Grupo suspeito de fornecer sementes de maconha para traficantes brasileiros é alvo de operação no DF e MG*. 2024. Website. Accessed: 2025-03-06. Disponível em: <https://g1.globo.com/df/distrito-federal/noticia/2024/02/27/gr upo- suspeito- for necer- sementes- de- maconha- para- traficantes- brasileiros- e- alvo- de- operacac no- df- e- mg.ghtml>. Citadonapágina21. G1. *Polícia Federal erradica 134 mil pés de maconha no Sertão de Pernambuco* 2025. Website. Accessed: 2025-03-06. Disponível em: <https://g1.globo.com/pe/petrolina-regiao/noticia/2025/02/24/policia-federal-erradica-134-mil-pes-de-maconha-no-sertao-de-pernambuco.ghtml>. Citado na página 21.

GAO, Y.; LI, S.; ZHAN, A. Genome-wide single nucleotide polymorphisms (snps) for a model invasive ascidian botryllus schlosseri. *Genetica*, Springer Science and Business Media LLC, v. 146, n. 2, p. 227–234, fev. 2018. ISSN 1573-6857. Disponível em:

<http://dx.doi.org/10.1007/s10709-018-0015-1>. Citadonapágina22.

GOUY, A.; ZIEGER, M. Straf—a convenient online tool for str data evaluation in forensic genetics. *Forensic Science International: Genetics*, Elsevier BV, v. 30, p. 148–151, set. 2017. ISSN 1872-4973. Disponível em: <http://dx.doi.org/10.1016/j.fsigen.2017.07.007>. Citado na página 83.

GRASSA, C. J. et al. A completecannabischromosome assembly and adaptive admixture for elevated cannabidiol (cbd) content. Cold Spring Harbor Laboratory, out. 2018. Disponível em: <http://dx.doi.org/10.1101/458083>. Citadonapágina80.

GREENER, J. G. et al. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, Springer Science and Business Media LLC, v. 23, n. 1, p. 40–55, set. 2021. ISSN 1471-0080. Disponível em: <http://dx.doi.org/10.1038/s41580-021-00407-0>. Citado na página 24.

GUO, C. et al. Transversions have larger regulatory effects than transitions. *BMC Genomics*, Springer Science and Business Media LLC, v. 18, n. 1, maio 2017. ISSN 1471-2164. Disponível em: <http://dx.doi.org/10.1186/s12864-017-3785-4>. Citado na página 22.

Health Canada. *Canadian Alcohol and Drugs Survey 2019: Summary*. 2019. Acesso em: 6 mar. 2025. Disponível em: <https://www.canada.ca/en/health-canada/services/canadian-alcohol-drugs-survey/2019-summary.html>. Citadonapágina19.

HOCH, E. et al. Cannabis, cannabinoids and health: a review of evidence on risks and medical benefits. *European Archives of Psychiatry and Clinical Neuroscience*, Springer Science and Business Media LLC, set. 2024. ISSN 1433-8491. Disponível em: <http://dx.doi.org/10.1007/s00406-024-01880-2>. Citado na página 19.

HUSSAIN, T. et al. Cannabis sativa research trends, challenges, and new-age perspectives. *iScience*, Elsevier BV, v. 24, n. 12, p. 103391, dez. 2021. ISSN 2589-0042. Disponível em: <http://dx.doi.org/10.1016/j.isci.2021.103391>. Citado2vezesnaspáginas13and17.

HUYNH, P.-H.; NGUYEN, V. H.; DO, T.-N. Improvements in the large p, small n classification issue. *SN Computer Science*, Springer Science and Business Media LLC, v. 1, n. 4, jun. 2020. ISSN 2661-8907. Disponível em: <http://dx.doi.org/10.1007/s42979-020-00210-2>. Citado na página 36.

IFTIKHAR, A. et al. Applications of cannabis sativa l. in food and its therapeutic potential: From a prohibited drug to a nutritional supplement. *Molecules*, MDPI AG, v. 26, n. 24, p. 7699, dez. 2021. ISSN 1420-3049. Disponível em: <http://dx.doi.org/10.3390/molecules26247699>. Citado na página 13.

JAGANNATHAN, R. Identification of psychoactive metabolites from cannabis sativa, its smoke, and other phytocannabinoids using machine learning and multivariate methods. *ACS Omega*, American Chemical Society (ACS), v. 5, n. 1, p. 281–295, jan. 2020. ISSN 2470-1343. Disponível em: <http://dx.doi.org/10.1021/acsomega.9b02663>. Citado na página 25.

JANISCHEVSKY, D. A form of hemp in wild areas of southeastern russia. *NG Černyševskogo Universiteta*, v. 2, p. 3–17, 1924. Citado na página 15.

JIN, D. et al. Classification of cannabis strains in the canadian market with discriminant analysis of principal components using genome-wide single nucleotide polymorphisms. *PLOS ONE*, Public Library of Science (PLoS), v. 16, n. 6, p. e0253387, jun. 2021. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0253387>. Citadonapágina23.

KIM, J. H.; KIM, M. S.; SEO, Y. W. Overexpression of a plant u-box gene tapub4 confers drought stress tolerance in arabidopsis thaliana. *Plant Physiology and Biochemistry*, Elsevier BV, v. 196, p. 596–607, mar. 2023. ISSN 0981-9428. Disponível em: <http://dx.doi.org/10.1016/j.plaphy.2023.02.001>. Citadonapágina86.

KODUR, V. K.; NASER, M. Z. Classifying bridges for the risk of fire hazard via competitive machine learning. *Advances in Bridge Engineering*, Springer Science and Business Media LLC, v. 2, n. 1, jan. 2021. ISSN 2662-5407. Disponível em: <http://dx.doi.org/10.1186/s43251-020-00027-2>. Citado na página 30.

KOSUGI, S.; OHASHI, Y. Interaction of the arabidopsis e2f and dp proteins confers their concomitant nuclear translocation and transactivation. *Plant Physiology*, Oxford University Press (OUP), v. 128, n. 3, p. 833–843, mar. 2002. ISSN 0032-0889. Disponível em: <http://dx.doi.org/10.1104/pp.010642>. Citado na página 85.

KOWALCZYK, M. et al. Molecular markers used in forensic genetics. *Medicine, Science and the Law*, SAGE Publications, v. 58, n. 4, p. 201–209, set. 2018. ISSN 2042-1818. Disponível em: <http://dx.doi.org/10.1177/0025802418803852>. Citadonapágina22.

LAPIERRE, ; MONTHONY, A. S.; TORKAMANEH, D. Genomics-based taxonomy to clarify cannabis classification. *Genome*, Canadian Science Publishing, v. 66, n. 8, p. 202–211, ago. 2023. ISSN 1480-3321. Disponível em: <http://dx.doi.org/10.1139/gen-2023-0005>. Citado na página 13.

LI, H. et al. The sequence alignment/map format and samtools. *Bioinformatics*, Oxford University Press (OUP), v. 25, n. 16, p. 2078–2079, jun. 2009. ISSN 1367-4803. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btp352>. Citadonapágina80.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. *RITA - Revista de Informática Teórica e Aplicada*, XIV, n. 2, 2007. Citado 3 vezes nas páginas 24, 25, and 29.

LU, Y. et al. Hyperspectral imaging with machine learning to differentiate cultivars, growth stages, flowers, and leaves of industrial hemp (cannabis sativa l.). *Frontiers in Plant Science*, Frontiers Media SA, v. 12, fev. 2022. ISSN 1664-462X. Disponível em: <http://dx.doi.org/10.3389/fpls.2021.810113>. Citadonapágina25.

MALABADI, R. B. et al. Cannabis sativa: Difference between medical cannabis (maijuana or drug type) and industrial hemp. *GSC Biological and Pharmaceutical Sciences*, GSC Online Press, v. 24, n. 3, p. 377–381, set. 2023. ISSN 2581-3250. Disponível em: <http://dx.doi.org/10.30574/gscbps.2023.24.3.0393>. Citadonapágina16.

MALINDA, R. R. Biological data studies, scale-up the potential with machine learning. *European Journal of Human Genetics*, Springer Science and Business Media LLC, v. 31, n. 6, p. 619–620, abr. 2023. ISSN 1476-5438. Disponível em: <http://dx.doi.org/10.1038/s41431-023-01361-5>. Citado na página 25.

MCKENNA, A. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, Cold Spring Harbor Laboratory, v. 20, n. 9, p. 1297–1303, jul. 2010. ISSN 1088-9051. Disponível em: <http://dx.doi.org/10.1101/gr.107524.110>. Citadonapágina80.

MCPARTLAND, J. M. Cannabissystematics at the levels of family, genus, and species. *Cannabis and Cannabinoid Research*, Mary Ann Liebert Inc, v. 3, n. 1, p. 203–212, out. 2018. ISSN 2378-8763. Disponível em: <http://dx.doi.org/10.1089/can.2018.0039>. Citado na página 15.

MCPARTLAND, J. M.; GUY, G. W. Models of cannabis taxonomy, cultural bias, and conflicts between scientific and vernacular names. *The Botanical Review*, Springer Science and Business Media LLC, v. 83, n. 4, p. 327–381, jun. 2017. ISSN 1874-9372. Disponível em: <http://dx.doi.org/10.1007/s12229-017-9187-0>. Citadonapágina15.

MCPARTLAND, J. M.; GUY, G. W.; HEGMAN, W. Cannabis is indigenous to europe and cultivation began during the copper or bronze age: a probabilistic synthesis of fossil pollen studies. *Vegetation History and Archaeobotany*, Springer Science and Business Media LLC, v. 27, n. 4, p. 635–648, maio 2018. ISSN 1617-6278. Disponível em: <http://dx.doi.org/10.1007/s00334-018-0678-7>. Citadonapágina14.

MCPARTLAND, J. M.; HEGMAN, W.; LONG, T. Cannabis in asia: its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies. *Vegetation History and Archaeobotany*, Springer Science and Business Media LLC, v. 28, n. 6, p. 691–702, maio 2019. ISSN 1617-6278. Disponível em: <http://dx.doi.org/10.1007/s00334-019-00731-8>. Citado na página 13. MD, V. et al. *Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems*. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1907.12931>. Citado na página 80.

MEAH, F. et al. The effects of cannabis and cannabinoids on the endocrine system. *Reviews in Endocrine and Metabolic Disorders*, Springer Science and Business Media LLC, v. 23, n. 3, p. 401–420, ago. 2021. ISSN 1573-2606. Disponível em: <http://dx.doi.org/10.1007/s11154-021-09682-w>. Citadonapágina17.

MEI, B.; WANG, Z. An efficient method to handle the 'large p, small n' problem for genomewide association studies using haseman–elston regression. *Journal of Genetics*, Springer Science and Business Media LLC, v. 95, n. 4, p. 847–852, dez. 2016. ISSN 0973-7731. Disponível em: <http://dx.doi.org/10.1007/s12041-016-0705-3>. Citadonapágina36.

MERLIN, M. D. Man and marijuana. *l972). Fairleigh Dickin-son University Press, Rutherford-Madison-Teaneck, New Jersey*, 1972. Citado na página 13.

Ministério da Justiça e Segurança Pública. *II Relatório Brasileiro sobre Drogas: Sumário Executivo*. 2021. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/sua- protecao/politicas- sobre- drogas/arquivo- manual- de- avaliacao- e- alienacao- de- bens/SumarioExecutivoIIRelatrioBrasileirosobreDrogas.pdf>. Acesso em: 6 mar. 2025. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/sua-protecao/politicas- sobre- drogas/arquivo- manual- de- avaliacao- e- alienacao- de- bens/SumarioExecutivoIIRelatrioBrasileirosobreDrogas.pdf>. Citado2vezesnaspáginas 19 and 20.

Ministério das Comunicações. *PRF registra recorde histórico de apreensão de drogas nas rodovias federais*. 2025. Website. Accessed: 2025-03-06. Disponível em: <https://www.gov.br/secom/pt-br/assuntos/noticias/2025/janeiro/prf-registra-recorde-historico-de-apreensao-de-drogas-nas-rodovias-federais>. Citadonapágina 21.

MUROVEC, J. et al. Analysis of morphological traits, cannabinoid profiles, thcas gene sequences, and photosynthesis in wide and narrow leaflet high-cannabidiol breeding populations of medical cannabis. *Frontiers in Plant Science*, Frontiers Media SA, v. 13, fev. 2022. ISSN 1664-462X. Disponível em: <http://dx.doi.org/10.3389/fpls.2022.786161>. Citado 2 vezes nas páginas 16 and 17.

NETO, P. A. P.; PIERRO, L. M. C.; FERNANDES, S. T. Cannabis: 12, 000 years of experiences and prejudices. *Brazilian Journal Of Pain*, v. 6, n. Supl.2, 2023. ISSN 2595-0118. Disponível em: <http://dx.doi.org/10.5935/2595-0118.20230055-en>. Citadonapágina14.

NUNZIO, M. D.; BARROT-FEIXAT, C.; GANGITANO, D. Cannabis chloroplast snp data for biogeographic origin and crop-type determination on european samples. Elsevier BV, 2023. Disponível em: <http://dx.doi.org/10.2139/ssrn.4412715>. Citado na página 23.

NUNZIO, M. D.; BARROT-FEIXAT, C.; GANGITANO, D. Characterization and evaluation of nine cannabis sativa chloroplast snp markers for crop type determination and biogeographical origin on european samples. *Forensic Science International: Genetics*, Elsevier BV, v. 68, p. 102971, jan. 2024. ISSN 1872-4973. Disponível em: <http://dx.doi.org/10.1016/j.fsigen.2023.102971>. Citado 2 vezes nas páginas 22 and 23.

ONOFRI, C.; MANDOLINO, G. *Genomics and Molecular Markers in Cannabis sativa L.* Cham: Springer International Publishing, 2017. 319–342 p. ISBN 978-3-319-54564-6. Disponível em: <https://doi.org/10.1007/978-3-319-54564-6_15>. Citado na página 22.

PACIFICO, D. et al. Time course of cannabinoid accumulation and chemotype development during the growth of cannabis sativa l. *Euphytica*, Springer Science and Business Media LLC, v. 160, n. 2, p. 231–240, ago. 2007. ISSN 1573-5060. Disponível em: <http://dx.doi.org/10.1007/s10681-007-9543-y>. Citadonapágina17.

PISANTI, S.; BIFULCO, M. Medical cannabis: A plurimillennial history of an evergreen. *Journal of Cellular Physiology*, Wiley, v. 234, n. 6, p. 8342–8351, nov. 2018. ISSN 1097-4652. Disponível em: <http://dx.doi.org/10.1002/jcp.27725>. Citado 2 vezes nas páginas 13 and 14.

Polícia Federal. *Operação Semilha combate tráfico de sementes entre Uruguai e Brasil*. 2022. Website. Accessed: 2025-03-06. Disponível em: <https://www.gov.br/pf/pt-br/assuntos/noticias/2022/02/operacao-semilha-combate-trafico-de-sementes-entre-uruguai-e-brasil>. Citadona página 21.

Polícia Federal. *PF encerra oitava etapa de operação para erradicação de plantios de maconha no Brasil*. 2024. Website. Accessed: 2025-03-06. Disponível em: <https://www.gov.br/pf/pt-br/assuntos/noticias/2024/09/pf- encer ra- oitava- etapa- de- operacao- para- er radicacao- de- plantios- de- maconha- no- brasil> . Citado na página 21.

POZO, J. C. del et al. The balance between cell division and endoreplication depends on e2fc-dpb, transcription factors regulated by the ubiquitin-scfskp2a pathway inarabidopsis. *The Plant Cell*, Oxford University Press (OUP), v. 18, n. 9, p. 2224–2235, ago. 2006. ISSN 1532-298X. Disponível em: <http://dx.doi.org/10.1105/tpc.105.039651>. Citado na página 85.

POZO, J. c. del et al. The e2fc-dpb transcription factor controls cell division, endoreplication and lateral root formation in a scfskp2a- dependent manner. *Plant Signaling amp; Behavior*, Informa UK Limited, v. 2, n. 4, p. 273–274, jul. 2007. ISSN 1559-2324. Disponível em: <http://dx.doi.org/10.4161/psb.2.4.3897>. Citadonapágina85.

PUNJA, Z. K.; SUTTON, D. B.; KIM, T. Glandular trichome development, morphology, and maturation are influenced by plant age and genotype in high thc-containing cannabis (cannabis sativa l.) inflorescences. *Journal of Cannabis Research*, Springer Science and Business Media LLC, v. 5, n. 1, abr. 2023. ISSN 2522-5782. Disponível em: <http://dx.doi.org/10.1186/s42238-023-00178-9>. Citado na página 17.

RADWAN, M. M. et al. *Natural Cannabinoids of Cannabis and Methods of Analysis*. Cham: Springer International Publishing, 2017. 161–182 p. Disponível em: <https://doi.org/10.1007/978-3-319-54564-6_7>. Citadonapágina17.

RAIHAN, A.; BIJOY, T. R. A review of the industrial use and global sustainability of cannabis sativa. *Global Sustainability Research*, Global Scientific Research, v. 2, n. 4, p. 1–29, out. 2023. ISSN 2833-986X. Disponível em: <http://dx.doi.org/10.56556/gssr.v2i4.597>. Citado na página 16.

RANSING, R. et al. Current state of cannabis use, policies, and research across sixteen countries: cross-country comparisons and international perspectives. *Trends in Psychiatry and Psychotherapy*, EDITORA SCIENTIFIC, 2021. ISSN 2238-0019. Disponível em: <http://dx.doi.org/10.47626/2237-6089-2021-0263>. Citadonapágina19.

RAPIN, L. et al. Cannabidiol use and effectiveness: real-world evidence from a canadian medical cannabis clinic. *Journal of Cannabis Research*, Springer Science and Business Media LLC, v. 3, n. 1, jun. 2021. ISSN 2522-5782. Disponível em: <http://dx.doi.org/10.1186/s42238-021-00078-w>. Citado na página 19.

REEL, P. S. et al. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, Elsevier BV, v. 49, p. 107739, jul. 2021. ISSN 0734-9750. Disponível em: <http://dx.doi.org/10.1016/j.biotechadv.2021.107739>. Citado na página 25.

RIBEIRO, L. de O. P. et al. Evaluation of two 13-loci str multiplex system regarding identification and origin discrimination of brazilian cannabis sativa samples. *International Journal of Legal Medicine*, Springer Science and Business Media LLC, v. 134, n. 5, p. 1603–1612, jun. 2020. ISSN 1437-1596. Disponível em: <http://dx.doi.org/10.1007/s00414-020-02338-5>. Citado 2 vezes nas páginas 20 and 22.

ROCK, E. M.; PARKER, L. A. Constituents of cannabis sativa. In: . *Cannabinoids and Neuropsychiatric Disorders*. Springer International Publishing, 2020. p. 1–13. ISBN 9783030573690. Disponível em: <http://dx.doi.org/10.1007/978-3-030-57369-0_1>. Citado na página 17.

RULL, V. Origin, early expansion, domestication and anthropogenic diffusion of cannabis, with emphasis on europe and the iberian peninsula. *Perspectives in Plant Ecology, Evolution and Systematics*, Elsevier BV, v. 55, p. 125670, jun. 2022. ISSN 1433-8319. Disponível em: <http://dx.doi.org/10.1016/j.ppees.2022.125670>. Citado na página 14.

SEIJO-PARDO, B. et al. Biases in feature selection with missing data. *Neurocomputing*, Elsevier BV, v. 342, p. 97–112, maio 2019. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2018.10.085>. Citadonapágina82.

SHAHBAZI, F. et al. Cannabinoids and cannabinoid receptors: The story so far. *iScience*, Elsevier BV, v. 23, n. 7, p. 101301, jul. 2020. ISSN 2589-0042. Disponível em: <http://dx.doi.org/10.1016/j.isci.2020.101301>. Citado na página 18.

SILVA, D. M. d.; ROSA, A. F. d. A lei de drogas nº 11.343 de 2006 e a política de encarceramento em massa: O fomento para o crime organizado. In: PROGRAMA DE PóS-GRADUAçãO EM CIêNCIAS CRIMINAIS (PPGCCRIM), PONTIFíCIA UNIVERSIDADE CATóLICA DO RIO GRANDE DO SUL (PUCRS). *Anais do 11º Congresso Internacional de Ciências Criminais*. 2020. Acesso em: 6 mar. 2025. Disponível em: <https://repositorio.pucrs.br/dspace/handle/10923/15716>. Citadonapágina20.

SMALL, E.; CRONQUIST, A. A practical and natural taxonomy for cannabis. *Taxon*, JSTOR, p. 405–435, 1976. Citado 2 vezes nas páginas 15 and 16.

STAGINNUS, C.; ZöRNTLEIN, S.; MEIJER, E. de. A <scp>pcr</scp> marker linked to a <scp>thca</scp> synthase polymorphism is a reliable tool to discriminate potentially <scp>thc</scp>-rich plants of <scp>c</scp>annabis sativa l. *Journal of Forensic Sciences*, Wiley, v. 59, n. 4, p. 919–926, mar. 2014. ISSN 1556-4029. Disponível em: <http://dx.doi.org/10.1111/1556-4029.12448>. Citadonapágina17.

STELLA, N. THC and CBD: Similarities and differences between siblings. *Neuron*, Elsevier BV, v. 111, n. 3, p. 302–327, fev. 2023. Citado na página 19.

SUNDAR, G. V. H. et al. Plant histone variants at the nexus of chromatin readouts, stress and development. *Biochimica et Biophysica Acta (BBA) - General Subjects*, Elsevier BV, v. 1868, n. 2, p. 130539, fev. 2024. ISSN 0304-4165. Disponível em: <http://dx.doi.org/10.1016/j.bbagen.2023.130539>. Citadonapágina87.

TAHIR, M. N. et al. The biosynthesis of the cannabinoids. *Journal of Cannabis Research*, Springer Science and Business Media LLC, v. 3, n. 1, mar. 2021. ISSN 2522-5782. Disponível em: <http://dx.doi.org/10.1186/s42238-021-00062-4>. Citado 2 vezes nas páginas 17 and 18.

TRENNER, J. et al. Evolution and functions of plant u-box proteins: From protein quality control to signaling. *Annual Review of Plant Biology*, Annual Reviews, v. 73, n. 1, p. 93–121, maio 2022. ISSN 1545-2123. Disponível em: <http://dx.doi.org/10.1146/annurev-arplant-102720-012310>. Citado na página 86.

United Nations Office on Drugs and Crime. *Recommended Methods for the Identification and Analysis of Cannabis and Cannabis Products*. Vienna: United Nations Office on Drugs and Crime, 2022. Accessed: 2025-02-27. Disponível em: <https://www.unodc.org/documents/scientific/STNAR40_Rev2_Cannabis_FINAL.pdf>. Citado2vezesnaspáginas17and21.

United Nations Office on Drugs and Crime (UNODC). *World Drug Report 2024*. 2024. Accessed: 2025-03-06. Disponível em: <https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html>. Citadonapágina19.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995. ISBN978-1-4612-5721-4. Citadonapágina29.

VAVILOV, N. I. -... [S.l.]: .-. , 1922. Citado na página 15.

WATTS, G. Cannabis confusions. *BMJ*, BMJ, v. 332, n. 7534, p. 175–176, jan. 2006. ISSN 1468-5833. Disponível em: <http://dx.doi.org/10.1136/bmj.332.7534.175>. Citado na página 15.

World Health Organization. *Cannabis*. 2025. Disponível em: <https://www.who.int/teams/mental- health- and- substance- use/alcohol- dr ugs- and- addictive- behaviours/dr ugs- psychoactive/cannabis>. Acesso em: 6 mar. 2025. Disponível em: <https://www.who.int/teams/mental- health- and- substance- use/alcohol- dr ugs- and- addictive- behaviours/dr ugs- psychoactive/cannabis>. Citado na página 19.

XIA, R. et al. Development and validation of a novel and fast detection method for cannabis sativa: A 19-plex short tandem repeat typing system. *Frontiers in Plant Science*, Frontiers Media SA, v. 13, fev. 2022. ISSN 1664-462X. Disponível em: <http://dx.doi.org/10.3389/fpls.2022.837945>. Citado na página 22.

XIE, Z. et al. Cannabis sativa: origin and history, glandular trichome development, and cannabinoid biosynthesis. *Horticulture Research*, Oxford University Press (OUP), v. 10, n. 9, jul. 2023. ISSN 2052-7276. Disponível em: <http://dx.doi.org/10.1093/hr/uhad150>. Citado na página 13.

XU, C.; JACKSON, S. A. Machine learning and complex biological data. *Genome Biology*, Springer Science and Business Media LLC, v. 20, n. 1, abr. 2019. ISSN 1474-760X. Disponível em: <http://dx.doi.org/10.1186/s13059-019-1689-0>. Citado na página 24.

YANG, S. et al. Classification and localization of maize leaf spot disease based on weakly supervised learning. *Frontiers in Plant Science*, Frontiers Media SA, v. 14, maio 2023. ISSN 1664-462X. Disponível em: <http://dx.doi.org/10.3389/fpls.2023.1128399>. Citado na página 25.

ZHANG, Z. et al. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning. *Scientific Reports*, Springer Science and Business Media LLC, v. 8, n. 1, jul. 2018. ISSN 2045-2322. Disponível em:

<http://dx.doi.org/10.1038/s41598-018-28840-w>. Citado na página 29.

ZHOU, L. et al. Erratum to “phenotypic analysis of diseased plant leaves using supervised and weakly supervised deep learning”. *Plant Phenomics*, Elsevier BV, v. 5, p. 0033, 2023. ISSN 2643-6515. Disponível em: <http://dx.doi.org/10.34133/plantphenomics.0033>. Citado na página 25.

ZUARDI, A. W. History of cannabis as a medicine: a review. *Revista Brasileira de Psiquiatria*, EDITORA SCIENTIFIC, v. 28, n. 2, p. 153–157, jun. 2006. ISSN 1516-4446. Disponível em: <http://dx.doi.org/10.1590/S1516-44462006000200015>. Citadonapágina14.