WSTAR

II WORKSHOP ON STATISTICAL TOOLS AND ANALYSIS FOR SCIENTIFIC RESEARCH

2025

HOME EDITORA

II Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR)

"Acreditamos que um mundo melhor se faz com a difusão do conhecimento científico".

Equipe Home Editora

Renius Mello
Paulo Santana Pacheco
Jeriel Dias
Deborah Vasconcellos
José Dilson Francisco da Silva
Alexandre José Cichoski
Paulo Cezar Bastianello Campagnol
Fernando Miranda de Vargas Júnior
Cesar Henrique Espírito Candal Poli
(Organizadores)

II Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR)

Belém-PA
Home Editora
2025

Home Editora
CNPJ: 39.242.488/0002-80
www.homeeditora.com
contato@homeeditora.com
91988165332
Tv. Quintino Bocaiúva, 23011 - Batista Campos, Belém - PA, 66045-315

Editor-Chefe
Prof. Dr. Ednilson Ramalho
Revisão
Organizadores

Bibliotecária
Janaina Karina Alves Trigo Ramos
CRB-8/009166
Produtor editorial
Laiane Borges

Dados Internacionais de Catalogação na publicação (CIP)

M528

II Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR) / Renius Mello, Paulo Santana Pacheco, Jeriel Dias, Deborah Vasconcellos, José Dilson Francisco da Silva, Alexandre José Cichoski, Paulo Cezar Bastianello Campagnol, Fernando Miranda de Vargas Júnior, Cesar Henrique Espírito Candal Poli (Organizadores). – Belém: Home, 2025.

Livro digital; 90 p.

ISBN 978-65-6089-439-6
DOI 10.46898/home.6db43725-9f06-4b4c-98db-a906c7b84ae6

1. Estatística. 2. Pesquisa científica. 3. Ciência de dados. I. Mello, Renius et al. (Organizador). II. Título.

CDD 519.5

Índice para catálogo sistemático:
I. Estatística aplicada à pesquisa científica
II. Ciência de dados
III. Pesquisa científica – Metodologia

WSTAR

WORKSHOP ON STATISTICAL TOOLS
AND ANALYSIS FOR SCIENTIFIC RESEARCH

https://www.ufsm.br/eventos/wstar

Comissão organizadora:

Alexandre José Cichoski
Alice Veleda Wendt
Cesar Henrique Espírito Candal Poli
Deborah Vasconcellos
Fernando Miranda de Vargas Júnior
Jeriel Dias
José Dilson Francisco da Silva
Paulo Cezar Bastianello Campagnol
Paulo Santana Pacheco
Renius Mello

Capa: Jeriel Dias

Diagramação: Paulo Santana Pacheco e Renius Mello

Realização:

PPGCTA

Apoio:

UFSM

Centro de Ciências Rurais UFSM

SISPEC CYTED

AlProSos

Departamento de Tecnologia e Ciência dos Alimentos

UFSM MEATGROUP

Departamento de Zootecnia UFSM

SUMÁRIO

APRESENTAÇÃO

A análise estatística ocupa um papel central no avanço da ciência e no apoio a decisões estratégicas em diferentes setores da sociedade. Mais do que lidar com números, trata-se de transformar dados em conhecimento, capaz de gerar impacto real em pesquisa, inovação e desenvolvimento tecnológico.

Com esse propósito, a Universidade Federal de Santa Maria promove a **2ª edição do Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR 2025)**, consolidando-se como um espaço de integração entre a academia, a indústria e a sociedade. O evento busca capacitar acadêmicos, pesquisadores e profissionais no uso de ferramentas estatísticas e de ciência de dados aplicadas à solução de problemas complexos em múltiplas áreas do conhecimento.

A programação desta edição reúne especialistas renomados do Brasil e do exterior, trazendo palestras que vão desde o *data storytelling* e o uso do *SAS® Studio™* na ciência e tecnologia dos alimentos, até temas de fronteira como a arquitetura moderna da Plataforma *Viya™ 4* e a geração de dados biológicos sintéticos por modelos generativos. Trata-se de uma oportunidade única de explorar conceitos, técnicas e ferramentas que estão moldando o presente e o futuro da análise de dados.

Ao oferecer um espaço de troca de experiências, aprendizado prático e diálogo interdisciplinar, o WSTAR 2025 reafirma seu compromisso em formar profissionais mais preparados para enfrentar os desafios contemporâneos, atuando de forma crítica, inovadora e colaborativa.

Contamos com a sua participação e desejamos um bom evento a todos.

Comissão organizadora

Capítulo I

DATA STORYTELLING: CONECTANDO DADOS E DECISÕES COM O SAS VISUAL ANALYTICS

Isabela Gilho Teixeira Francisco[1]

DOI: 10.46898/home.9786560893306.1

1 SAS Brasil. Pós-Graduanda do Curso de Engenharia e Administração de Banco de Dados, Universidade Estadual de Campinas (UNICAMP), São Paulo, SP, e-mail: isabelagilhot@gmail.com

Resumo: *Data storytelling* é a disciplina que combina análise de dados, narrativa e design visual para transformar *insights* complexos em comunicações persuasivas que orientam decisões organizacionais. Este capítulo apresenta os fundamentos conceituais do *data storytelling,* seus elementos estruturais e metodologias. São apresentadas orientações para estruturar histórias orientadas a objetivos, considerando aspectos como clareza, contexto, público e impacto. O SAS Visual Analytics é apresentado como plataforma tecnológica que operacionaliza esses conceitos através de recursos de exploração aumentada, explicações automatizadas, objetos analíticos avançados e distribuição multiplataforma, facilitando a criação de narrativas visuais governadas e reprodutíveis para diferentes públicos e contextos.

Abstract: Data storytelling is the discipline that combines data analysis, narrative, and visual design to transform complex insights into persuasive communications that guide organizational decisions. This chapter introduces the conceptual foundations of data storytelling, its structural elements, and methodologies. It offers guidelines for crafting goal-oriented stories, addressing aspects such as clarity, context, audience, and impact. SAS Visual Analytics is presented as the technological platform that operationalizes these concepts through augmented exploration, automated explanations, advanced analytical objects, and multi-channel distribution, facilitating the creation of governed and reproducible visual narratives for diverse audiences and contexts.

1 INTRODUÇÃO

Data storytelling surge como uma abordagem essencial para conectar análises complexas ao processo de tomada de decisão, ultrapassando a simples apresentação de gráficos ou tabelas. Ao combinar dados confiáveis, uma sequência narrativa bem definida e *design* visual estratégico, o *data storytelling* transforma informações brutas em narrativas persuasivas que despertam interesse e orientam ações. O primeiro passo é contextualizar o problema de negócio e compreender o público-alvo, estabelecendo um fio condutor que guiará o leitor desde a identificação de padrões até a recomendação de soluções. Em seguida, selecionam-

se métricas e indicadores relevantes, preparando os dados por meio de regras de qualidade, variáveis derivadas e hierarquias que facilitem a exploração. Por fim, escolhem-se visualizações adequadas como, gráficos de tendência, diagramas de correlação, mapas ou indicadores de destaque, e organiza-se a narrativa em sequência lógica, com títulos conclusivos, anotações e realces que reforçam as mensagens-chave.

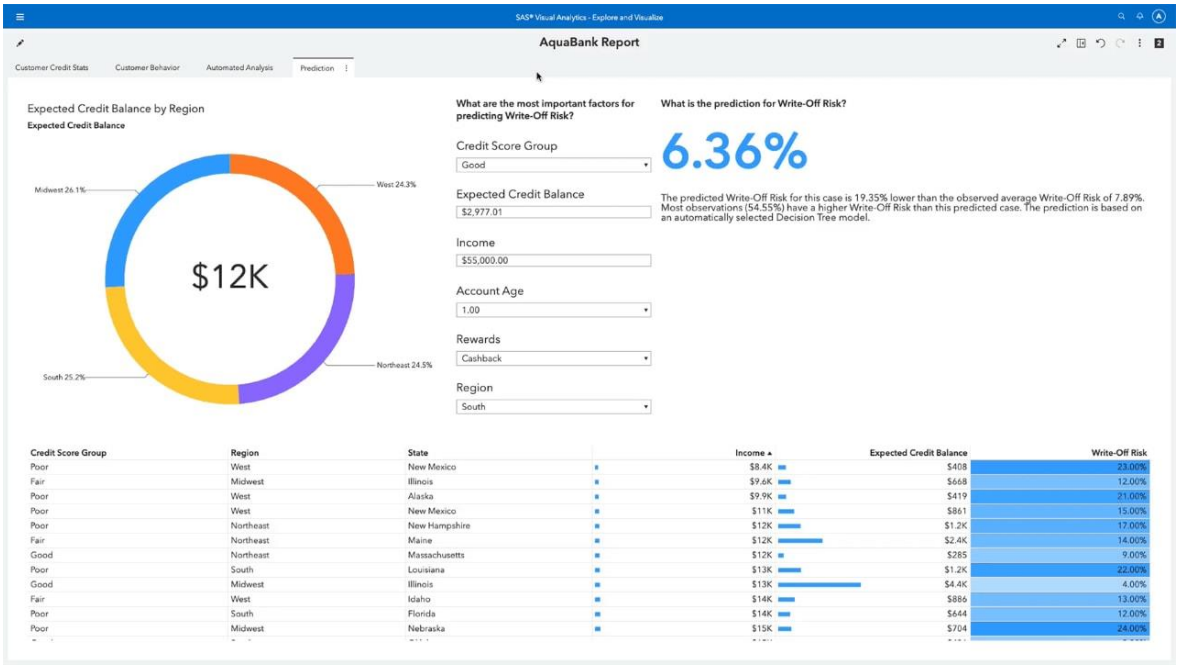
Além disso, o *data storytelling* exige atenção aos princípios de *design* e comunicação: a hierarquia visual deve conduzir o olhar do público, o uso de cores precisa ser consistente e as tipografias, legíveis. Anotações e texto dinâmico ajudam a situar datas, eventos ou filtros aplicados, enquanto as recomendações devem estar alinhadas a metas de negócio claras. A prática de contar histórias com dados depende também de uma metodologia iterativa: prototipar páginas, testar com usuários e ajustar elementos visuais e narrativos até que a mensagem seja transmitida de forma objetiva, evitando ruído cognitivo e garantindo que cada *insight* tenha impacto real.

Para apoiar nesse processo o SAS Institute, fundada em 1976, consolidou-se como referência global em soluções analíticas, abrangendo desde preparação e governança de dados até modelagem preditiva e inteligência artificial (IA). Entre as diversas soluções do ecossistema SAS, o SAS Visual Analytics (SVA), apresentado na Figura 1, desempenha um papel central na prática do *data storytelling*. Trata-se de uma ferramenta que permite explorar, visualizar e comunicar insights de maneira interativa e intuitiva, transformando dados em narrativas visuais impactantes. O SVA combina recursos de exploração, explicações automatizadas, objetos analíticos avançados e distribuição multiplataforma, facilitando a criação de painéis explicativos e histórias orientadas à ação. Dessa forma, a ferramenta não apenas viabiliza a análise, mas também aprimora a comunicação dos resultados, tornando o processo decisório mais ágil, transparente e colaborativo. Além disso, essa solução permite que profissionais de diferentes perfis construam, validem e distribuam narrativas visuais governadas e reproduzíveis, conectando dados a decisões de forma ágil e confiável.

A partir disso, este trabalho busca explicar de forma abrangente o conceito de *data storytelling*, abordando seus fundamentos teóricos e metodológicos,

detalhando as métricas e indicadores essenciais para estruturar narrativas analíticas eficazes e demonstrando como o SAS, por meio do SAS Visual Analytics, disponibiliza um conjunto integrado de recursos desde a preparação e exploração de dados até objetos analíticos avançados e explicações automatizadas que viabilizam a construção de histórias visuais governadas, interativas e alinhadas às necessidades de decisão organizacional.

Figura 1: Demonstração de relatório no SAS Visual Analytics.



2 METODOLOGIAS

A elaboração de narrativas analíticas eficazes requer uma abordagem sistemática que una práticas de análise de dados, princípios de comunicação e técnicas de *design* visual. Neste capítulo são apresentadas as etapas metodológicas para estruturar e validar histórias de dados, incluindo a seleção e preparação das fontes, a definição do fluxo narrativo, a aplicação de conceitos de comunicação e o uso de padrões de visualização. Cada seção fornece orientações práticas e referências a ferramentas que suportam o trabalho, garantindo rigor e repetibilidade.

2.1 Seleção e Preparação de Dados

Antes de contar qualquer história com dados, é essencial reunir informações relevantes e torná-las confiáveis. Primeiro, identifique as fontes de dados necessárias, como sistemas internos, planilhas ou bancos de dados, e verifique se elas contêm as informações certas para responder às perguntas de negócio. Diferentes conjuntos de dados podem ser integrados para criar uma visão mais completa do problema ou oportunidade analisada, garantindo que os *insights* gerados reflitam a realidade do contexto investigado.

Em seguida, os dados passam por uma etapa de preparação, que envolve processos de limpeza, padronização e organização. Isso inclui corrigir formatos inconsistentes (como datas registradas de maneiras diferentes), tratar valores ausentes ou duplicados e estruturar variáveis de forma coerente. O objetivo é garantir que os dados sejam consistentes, confiáveis e adequados para a análise, reduzindo o risco de interpretações equivocadas e fortalecendo a credibilidade das narrativas construídas.

A preparação de dados também deve considerar governança e rastreabilidade, assegurando que cada transformação, cálculo ou integração de fonte seja documentado e auditável. Essa prática permite que os dados possam ser reutilizados de forma segura em diferentes contextos analíticos, mantendo integridade e transparência.

Ao garantir que os dados estejam confiáveis, estruturados e contextualizados, essa etapa fornece a base sólida necessária para o *data storytelling,* permitindo que *insights* complexos sejam comunicados de maneira clara, persuasiva e orientada a decisões.

2.2 Definição do Fluxo Narrativo

Com os dados estruturados, mapeia-se a sequência lógica que guiará o público do diagnóstico ao insight e, finalmente, à ação recomendada. O fluxo inclui introdução do problema, apresentação de evidências e contextualização dos resultados, culminando em conclusões claras. Cada segmento narrativo é associado a objetivos específicos, mensagens centrais e perguntas que devem ser

respondidas por gráficos ou explicações em linguagem natural. O planejamento do fluxo narrativo é essencial para guiar o leitor do entendimento do problema até a recomendação de forma fluida e impactante. Ao definir esse percurso, combinam-se técnicas clássicas de *storytelling* adaptadas ao contexto analítico, garantindo que cada etapa cumpra um propósito claro e mantenha o engajamento. A seguir, discutem-se os principais conceitos que norteiam essa construção.

No fluxo de narrativa clássica (Início, meio e fim) no início, apresenta-se o contexto de negócio e o desafio que motiva a análise, estabelecendo as perguntas centrais que orientarão toda a narrativa. No meio, desenvolve-se a exploração dos dados, revelando padrões, hipóteses e conflitos que surgem durante a investigação. No fim, expõem-se os principais achados e recomendações, traduzindo insights em ações práticas. Essa segmentação assegura clareza e faz com que o leitor saiba sempre em que etapa da história se encontra.

Já a narrativa da jornada do herói ajuda a posicionar o público como protagonista da narrativa. Primeiro, ocorre o “chamado à aventura”, que corresponde ao problema de negócio. Em seguida, a travessia de testes, representada pela análise exploratória e pela validação de hipóteses, traz conflitos, como dados inesperados ou discrepantes. A “recompensa” final é o insight principal que capacita a decisão. Essa analogia literária mantém o interesse e reforça a ideia de superação de desafios.

2.3 Conceitos de Comunicação Analítica

As histórias de dados devem utilizar técnicas de redação e argumentação que facilitem a compreensão e estimulem a confiança do público. Isso envolve a escolha de títulos conclusivos, anotações explicativas, exemplos ilustrativos e texto dinâmico que reflita seleções de filtro ou períodos analisados. A consistência terminológica e a adaptação ao nível de conhecimento dos *stakeholders* são fundamentais para evitar ambiguidades e reforçar a credibilidade.

Durante a comunicação analíticas, uma das técnicas conhecidas e mais difundidas é o AMA (Anáfora, Metáfora e Analogia), utilizada para reforçar sua mensagem enriquecendo a comunicação ao empregar três recursos retóricos que facilitam o entendimento de conceitos e reforçam a mensagem central. A anáfora consiste em repetir palavras ou estruturas no início de frases para criar

ritmo e destaque, por exemplo, iniciando vários pontos com “Sempre que identificamos…”, o que reforça padrões e regras de negócio. A metáfora estabelece comparações diretas entre conceitos técnicos e situações cotidianas, como descrever um funil de vendas como “um coador que retém oportunidades até a etapa de fechamento”, aproximando o leitor de ideias abstratas. Já a analogia relaciona duas ideias distintas para explicar relações complexas, por exemplo, comparar um dashboard a um mapa de navegação que orienta decisões estratégicas ao mostrar rotas e destinos claros. Ao combinar anáfora, metáfora e analogia, o AMA torna as narrativas de dados mais envolventes, memoráveis e acessíveis, garantindo que cada insight seja compreendido e retenha seu impacto.

2.4 Visualização de Dados

A visualização é um componente essencial do *data storytelling*, funcionando como a ponte entre informação e compreensão, capaz de transformar dados complexos em representações visuais claras e interpretáveis. O uso adequado de cores, formas, tamanhos e *layouts* destaca padrões, comparações e tendências, facilitando a percepção rápida e correta pelo público. A escolha do tipo de gráfico, como barras, linhas, mapas ou diagramas de dispersão, deve ser cuidadosamente alinhada ao tipo de dado, à mensagem que se deseja transmitir e à complexidade do insight, pois nem todos os gráficos são adequados para todas as análises.

Além disso, a visualização deve seguir princípios de clareza e hierarquia visual, orientando o observador pelos elementos mais importantes da narrativa, e incorporar interatividade sempre que possível, permitindo explorar os dados de forma dinâmica e aprofundar a análise. A integração de *storytelling* visual com elementos de *design*, como destaques, agrupamentos e sequências lógicas, potencializa a compreensão e o engajamento.

O SAS Visual Analytics oferece recursos avançados para suportar essas práticas, permitindo criar *dashboards* interativos e visualizações dinâmicas que combinam cores, gráficos e objetos analíticos de maneira intuitiva. Com recursos como exploração aumentada, explicações automáticas e filtros interativos, a plataforma facilita a transformação de dados em narrativas visuais impactantes, governadas e orientadas à tomada de decisão.

2.5 Avaliação, Iteração e Governança

Por fim, valida-se a história por meio de testes com usuários, revisões de *design* e checagens de acessibilidade. Ferramentas de revisão automática e políticas de governança garantem consistência de formato, controle de versões e rastreabilidade de alterações. Com base no *feedback*, iteram-se visualizações e narrativas até que a mensagem seja transmitida de maneira clara e impactante, assegurando que cada elemento contribua para os objetivos de decisão organizacional.

3 VISUALIZAÇÕES DE DADOS COM O SAS VISUAL ANALYTICS

O SAS Visual Analytics (SVA), solução integrada ao ecossistema SAS, vai além de simples gráficos ao unir poder analítico e visualização interativa em um único ambiente. Neste capítulo, serão explorados exemplos de objetos e relatórios que ilustram como essa ferramenta pode materializar os princípios do *data storytelling,* oferecendo *dashboards* e painéis que orientam o usuário desde a descoberta de padrões até a recomendação de ações, suportando a tomada de decisão com clareza e profundidade visual.

A visualização gráfica é a espinha dorsal de relatórios eficazes, pois traduz dados brutos em *insights* imediatos e intuitivos. Gráficos bem projetados facilitam a detecção de padrões, comparações e exceções, reduzindo o esforço cognitivo necessário para interpretar grandes volumes de informação. Ao orientar o olhar do usuário para os elementos críticos, sejam tendências temporais, distribuições geográficas ou relações entre variáveis, eles tornam a narrativa mais envolvente e direcionada à ação.

No SAS Visual Analytics, é possível criar esses gráficos com rapidez e precisão. A plataforma oferece uma biblioteca rica de tipos de visualização pré-configurados que podem ser personalizáveis em cores, rótulos e estilos, como pode ser visto na apresentação de objetos gráficos nas Figuras 2, 3 e 4. Essa combinação de flexibilidade, interatividade e governança faz do SAS Visual Analytics uma

ferramenta poderosa para construir relatórios visuais que capturam a atenção, esclarecem insights e orientam decisões.

O SAS disponibiliza uma galeria pública, encontrada no site SAS Visual Analytics Gallery com exemplos de relatórios, de onde foram extraídas as imagens seguintes, para demonstrar como visualizações bem elaboradas podem orientar efetivamente o *data storytelling*.

Figura 2: Visualizações gráficas a partir de objetos de tabela encontrados no SVA.

Supplier Name	Number of Products	Profit
3Top Sports	421	$752,811.83
A Team Sports	70	$225,763.55
A. Pereira Sport	1	$185.50
AllSeasons Outdoor Clothing	68	$807,612.30
Bianco Trading S.A.	4	$9,306.90
Bon Garrments	7	($1,028.90)
British Sports Ltd	42	$641,644.65
Carolina Sports	13	$3,798.81
CrystalClear Optics Inc	71	$286,423.94
Dolphin Sportswear Inc	20	$84,919.00
Dutchman Bikes	1	$23.60
EA Sports Limited	10	$24,048.92
Eclipse Inc	633	($141,369.20)

Air Quality Index (AQI) duration (number of days)

County	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous
Alameda	236	110	8	11	1	
Amador	327	29	2	1		
Butte	183	144	18	18	2	1
Calaveras	260	81	10	13		
Colusa	247	87	9	23		
Contra Costa	258	89	8	11		
El Dorado	256	85	17	6		
Fresno	94	173	59	34	6	
Glenn	251	84	7	17		
Humboldt	333	30	3			
Imperial	119	219	26	2		
Inyo	172	139	23	19	5	8
Kern	78	158	98	30	2	

Figura 3: Visualizações gráficas a partir de objetos gráficos encontrados no SVA.

Figura 4: Visualizações gráficas a partir de objetos geográficos encontrados no SVA.

4 RESULTADOS

Nesta seção, propõe-se uma análise orientada por *data storytelling* que conecta diretamente dados e decisões a partir do painel criado no SAS Visual Analytics, apresentado na Figura 5. Este relatório foi construído a partir de uma base pública chamada '*Home Equity'*, disponibilizada no site GitHub SAS Viya.

Figura 5: Relatório construído no SVA para análise.

Aplicando nossa metodologia de *data storytelling* para a construção do relatório acima, primeiro foi realizada a seleção e preparação dos dados ao integrar fontes

de empréstimos, limpar registros inconsistentes e criar indicadores chave. Em seguida, definiu-se o fluxo narrativo estruturado buscando trazer uma ordem de análise para os dados apresentados no relatório de forma gráfica. Na etapa de comunicação, empregamos títulos conclusivos e analogias que aproximam conceitos técnicos a cenários familiares, enquanto no design de visualizações aplicamos paletas de cores consistentes, hierarquia de elementos e interatividade para destacar tendências, que possibilitou gerar os seguintes *insights*:

- 80% dos empréstimos estão em conformidade e 20% em inadimplência, apontando para ações de cobrança segmentadas;
- A maioria dos empréstimos fica entre 10.000 e 30.000 unidades, definindo o *ticket* médio predominante e orientando políticas de crédito;
- Profissionais autônomos têm valor médio de empréstimo mais alto (cerca de 28.000), seguidos por gerentes e professores, indicando perfis para ofertas customizadas;
- O número de novos empréstimos caiu de aproximadamente 120 em 2017 para 70 em 2021, sinalizando necessidade de campanhas de estímulo;
- A dívida acumulada prévia não impede, isoladamente, a concessão de créditos mais elevados, sugerindo a importância de modelagem de risco multidimensional.

5 CONCLUSÕES

O *data storytelling* se consolidou como uma disciplina essencial para transformar dados em decisões estratégicas, unindo análise, narrativa e *design* visual. Ao estruturar informações de forma clara, contextualizada e persuasiva, profissionais e organizações conseguem comunicar *insights* complexos de maneira eficaz, aumentando a compreensão e a ação baseada em dados.

O SAS Visual Analytics surge como uma ferramenta que potencializa essa prática, oferecendo recursos de exploração interativa, visualizações dinâmicas e explicações automatizadas, que facilitam a construção de narrativas analíticas

consistentes, governadas e replicáveis. Dessa forma, a combinação de métodos sólidos de *data storytelling* com tecnologias avançadas permite que dados deixem de ser apenas números e se tornem histórias capazes de orientar decisões, gerar impacto e promover resultados significativos nas organizações.

6 AGRADECIMENTOS

Registro minha gratidão ao SAS Brasil e à Deborah Vasconcellos, cuja atuação como Sr. Global Academic Program Manager abriu caminhos para a execução deste trabalho e para a apresentação da tecnologia SAS ao público da Universidade Federal de Santa Maria (UFSM).

Agradeço ainda à WSTART (*Workshop on Statistical Tools and Analysis for Scientific Research*) da UFSM pela oportunidade de participação, bem como a toda equipe organizadora do evento, especialmente ao Renius Mello, pela atenção e profissionalismo na organização do workshop.

REFERÊNCIAS

KNAFLIC, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.

MURRAY, S. (2017). *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. O'Reilly Media.

SAS INSTITUTE. (2025). SAS Visual Analytics. Disponível em: https://www.sas.com/en_us/software/visual-analytics.html. Acesso em 08/10/2025.

SAS INSTITUTE. (2025). *Welcome to SAS Visual Analytics*. Disponível em: https://documentation.sas.com/doc/pt-BR/vacdc/v_034/vawlcm/home.htm. Acesso em 08/10/2025.

Capítulo II

USO DO NOVO SAS STUDIO NA CIÊNCIA E TECNOLOGIA DE ALIMENTOS

Sérgio da Costa Côrtes[1]

DOI: 10.46898/home.9786560893306.2

1 Instituto de Educação Superior de Brasília. Professor do Departamento de Ciência de Dados e Inteligência Artificial, IESB, Brasília, DF, sergio.cortes@iesb.edu.br. ID Lattes: 1206696860799261

Resumo: O minicurso *"Uso do novo SAS Studio na Ciência e Tecnologia de Alimentos"* tem como objetivo introduzir estudantes e profissionais da área ao uso prático da plataforma SAS Studio para análise e visualização de dados aplicados às Ciências dos Alimentos. Utilizando uma abordagem metodológica baseada nos Steps visuais, o curso guia os participantes pelas etapas essenciais de um fluxo analítico: desde a organização das pastas, criação de bibliotecas (LIBNAME) e importação de dados, até a execução de estatísticas descritivas, análises de frequências e criação de gráficos automatizados. Essa abordagem sem necessidade de codificação facilita a compreensão e a aplicação dos conceitos estatísticos fundamentais para o controle de qualidade, pesquisa e desenvolvimento de produtos alimentícios, promovendo a autonomia analítica e o uso de boas práticas na análise de dados laboratoriais, físico-químicos e sensoriais.

Palavras-chave: SAS Studio; Ciência de Alimentos; Estatística Descritiva; Visualização de Dados; Steps.

Abstract: The short course *"Using the New SAS Studio in Food Science and Technology"* is designed to introduce students and professionals to the practical use of the SAS Studio platform for data analysis and visualization in the field of Food Science. Following a structured, step-by-step methodology based on the platform's intuitive visual Steps, the course guides participants through the essential stages of a data analytics workflow: organizing work folders, creating permanent libraries using the LIBNAME statement, importing structured datasets, and performing descriptive statistics, frequency analysis, and automated visualizations. This no-code, point-and-click approach simplifies the application of statistical concepts, empowering learners to analyze laboratory, physicochemical, and sensory data effectively. By the end of the course, participants will be equipped with the analytical skills and good practices needed to support decision-making in quality control, product development, and scientific research in food-related contexts.

Keywords: SAS Studio; Food Science; Descriptive Statistics; Data Visualization; Steps.

1 INTRODUÇÃO

O **SAS** - originalmente um acrônimo para ***Statistical Analysis System*** — nasceu nos anos 1970 como um projeto acadêmico na *Universidade Estadual da Carolina do Norte (EUA)*, com o objetivo de desenvolver um sistema computacional para *análise estatística de dados agrícolas*. Desde então, evoluiu para se tornar uma das mais completas e respeitadas plataformas de **análise de dados, estatística avançada, inteligência artificial e ciência de dados** no mundo. Hoje, o **SAS** é amplamente utilizado em organizações globais líderes nos setores de saúde, finanças, governo, educação e tecnologia, oferecendo soluções que vão da análise preditiva à inteligência de negócios. No meio acadêmico, o **SAS** é referência no **ensino da estatística e da inteligência artificial aplicada**, preparando gerações de estudantes e profissionais para os desafios da economia orientada por dados.

O **SAS Viya**, plataforma em nuvem de última geração, representa a evolução do **SAS** para o contexto da **análise moderna, escalável e colaborativa**. Projetado para integrar *ciência de dados, aprendizado de máquina, inteligência artificial e gestão de dados* em um único ambiente, o Viya permite executar análises em grandes volumes de informação, com desempenho otimizado e suporte a múltiplas linguagens como **SAS, Python, R, Lua e SQL**. Sua arquitetura aberta e baseada em APIs facilita a integração com outras tecnologias, ao mesmo tempo em que mantém a robustez e confiabilidade que tornaram o **SAS** uma referência mundial. Essa flexibilidade torna o **Viya** uma plataforma estratégica tanto para organizações quanto para a formação acadêmica em áreas interdisciplinares.

O avanço da **Ciência e Tecnologia de Alimentos** depende cada vez mais do uso estratégico de dados para compreender fenômenos complexos, otimizar processos produtivos e garantir qualidade, segurança e inovação em produtos alimentícios. Nesse cenário, o **SAS Studio**, componente do **SAS Viya**, em sua versão mais recente, surge como uma poderosa plataforma para exploração, manipulação, análise estatística e visualização de dados, permitindo que estudantes, pesquisadores e profissionais transformem informações em conhecimento aplicável. Este minicurso tem como propósito apresentar os recursos do **novo SAS Studio** aplicados ao contexto da área de alimentos, explorando desde

a organização e importação de bases de dados até a geração de análises descritivas, inferenciais e relatórios gráficos interativos. A proposta é oferecer uma experiência prática, acessível e orientada a problemas reais da ***Ciência e Tecnologia de Alimentos***, demonstrando como a análise de dados pode apoiar pesquisas científicas, desenvolvimento de novos produtos e processos de controle de qualidade.

2 METODOLOGIA PARA ANÁLISE DE DADOS

A metodologia proposta para o ***Uso do novo SAS Studio na ciência e tecnologia de alimentos*** visa oferecer uma experiência prática e guiada no uso do **SAS Studio** para exploração e análise de dados na área de *Ciência e Tecnologia de Alimentos*. O enfoque está em conduzir os alunos, passo a passo, desde a organização inicial dos arquivos até a aplicação das ferramentas de estatística descritiva, utilizando os **Steps do SAS Studio** em ambiente gráfico (*point and click*). Essa abordagem permitirá que os alunos compreendam não apenas os conceitos técnicos, mas também o *fluxo de trabalho* recomendado para transformar dados brutos em informações úteis, integrando boas práticas de organização, importação e análise de dados.

1. **Organização Inicial dos Arquivos**

Antes do início das análises, os participantes serão orientados a **criar uma pasta específica no ambiente do SAS Studio (Server Files and Folders)** para armazenar todos os arquivos utilizados no exercício. Esse procedimento garante melhor organização, facilita a importação de dados e evita problemas de localização durante o minicurso.

2. **Criação de uma Library no SAS**

O próximo passo consiste na criação de uma **SAS Library permanente** que aponte para a pasta recém-criada. Por meio do comando LIBNAME ou pela interface gráfica de **New Library**, os participantes aprenderão a vincular dados externos ao ambiente SAS. Essa etapa é fundamental para estruturar o fluxo de trabalho, assegurando que os datasets possam ser reutilizados em análises futuras.

3. **Importação dos Arquivos**

Na sequência, será utilizado o **Step "Import Data"** para carregar arquivos em formatos como **CSV** ou **Excel** para dentro da biblioteca criada. O processo inclui:

- Seleção do arquivo de origem na pasta do aluno.
- Definição da biblioteca de destino e nome da tabela SAS.
- Ajustes de delimitadores, nomes de variáveis e tipos de dados.

Essa etapa será feita integralmente em ambiente gráfico (*point and click*), sem necessidade de programação.

4. **Exploração Inicial dos Dados**

Após a importação, os alunos realizarão uma inspeção dos *datasets* importados:

- Visualização do conteúdo das tabelas.
- Verificação do número de observações e variáveis.
- Identificação de variáveis numéricas e categóricas.

Essas ações visam familiarizar o aluno com a base de dados a ser analisada.

5. **Análises Descritivas com Steps**

Com os dados já preparados, os participantes aplicarão os **Steps** de ***estatística*** e visualização:

- **Summary Statistics** (PROC MEANS): cálculo de média, mediana, mínimo, máximo, desvio padrão e contagem.
- **Distribution Analysis** (PROC UNIVARIATE): análise da distribuição das variáveis numéricas, histogramas e medidas de assimetria e curtose.
- **Frequency Analysis** (PROC FREQ): geração de tabelas de frequência para variáveis categóricas.

6. **Síntese e Discussão**

Por fim, os resultados obtidos serão discutidos coletivamente, destacando como os recursos do SAS Studio podem apoiar a **exploração de dados na área de**

Ciência e Tecnologia de Alimentos, permitindo identificar padrões, tendências e características importantes dos conjuntos de dados.

3 BASE DE DADOS

Para o desenvolvimento desta atividade usaremos a base de dados "*Worldwide Meat Consumption*", extraída de https://www.kaggle.com/datasets/vagifa/meatconsumption[3]. Esta base de dados está assim contextualizada:

"*O consumo de carne está relacionado aos padrões de vida, à dieta alimentar, à produção pecuária e aos preços ao consumidor, bem como à incerteza macroeconômica e aos choques no PIB. Comparada a outras commodities, a carne é caracterizada por altos custos de produção e altos preços de venda. A demanda por carne está associada a rendas mais altas e a uma mudança – devido à urbanização – no consumo de alimentos que favorecem o aumento da proteína de origem animal na dieta. Embora a indústria global da carne forneça alimento e sustento para bilhões de pessoas, ela também tem consequências ambientais e de saúde significativas para o planeta.*

Este conjunto de dados foi atualizado em 2018, com projeções mundiais para a carne até 2026 apresentadas para carne bovina e de vitelo, suína, de aves e ovina. O consumo de carne é medido em milhares de toneladas de peso da carcaça (exceto para aves, expresso em peso pronto para cozinhar) e em quilogramas de peso no varejo per capita. Os fatores de conversão do peso da carcaça para o peso no varejo são: 0,7 para carne bovina e de vitelo, 0,78 para carne suína e 0,88 para carne ovina e de aves. Exclui a Islândia, mas inclui todos os 28 países membros da União Europeia".

Conteúdo

O arquivo **.csv** possui 5 colunas:

- LOCATION = nome do código do país
- SUBJECT = tipo de carne (suíno, bovino, etc.)

[3] Acessada em 05/10/2025

- TIME = ano em que os dados foram registrados
- MEASURE = medida usada para mostrar o valor
- VALUE = valor, de acordo com a medida

Conteúdo

As unidades de medidas usadas para mensurar o consumo de carnes são as seguintes:

- **KG_CAP**

  A medida **KG_CAP** significa "**quilogramas por habitante**" - ou, mais precisamente, *quilogramas per capita*.

  Interpretação:

  - **KG** → quilogramas (unidade de massa no Sistema Internacional, equivalente a 1.000 gramas);
  - **CAP** → abreviação de *capita*, do latim, que significa "por pessoa".

  Portanto:

  1 KG_CAP = 1 quilograma por habitante**.**

- **THND_T.**

  A medida **THND_T (THND_TONNE)** significa literalmente "**thousand tonnes**", ou em português, "mil toneladas".

  Interpretação:

  - THND → abreviação de *thousand* (mil);
  - TONNE → unidade métrica de massa equivalente a 1.000 kg (também chamada de "tonelada métrica").

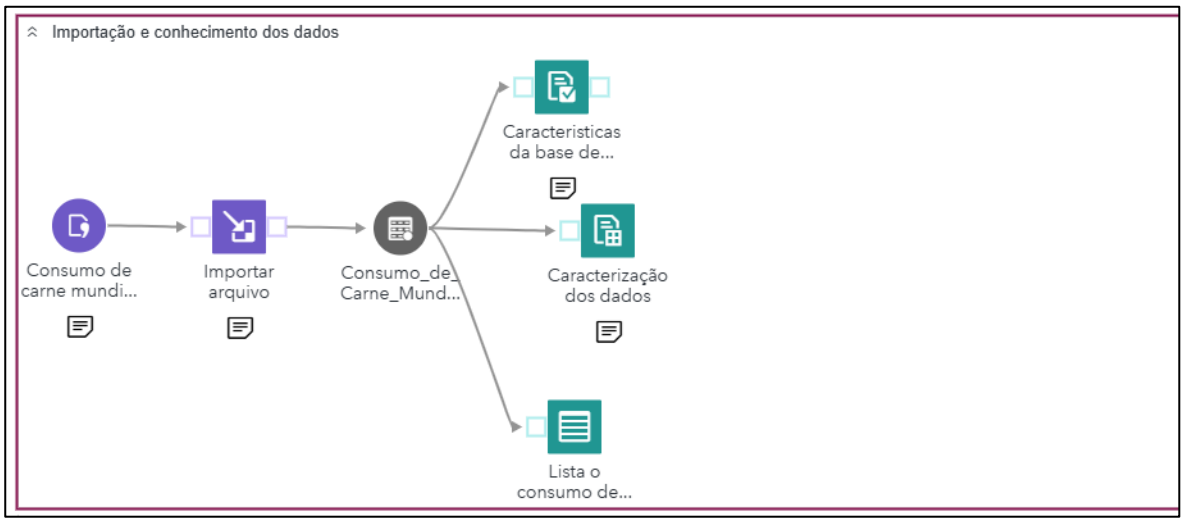
  Portanto:

  1 THND_TONNE = 1.000 toneladas métricas=1.000.000 kg.

Aplicação da Metodologia e uso do SAS Studio

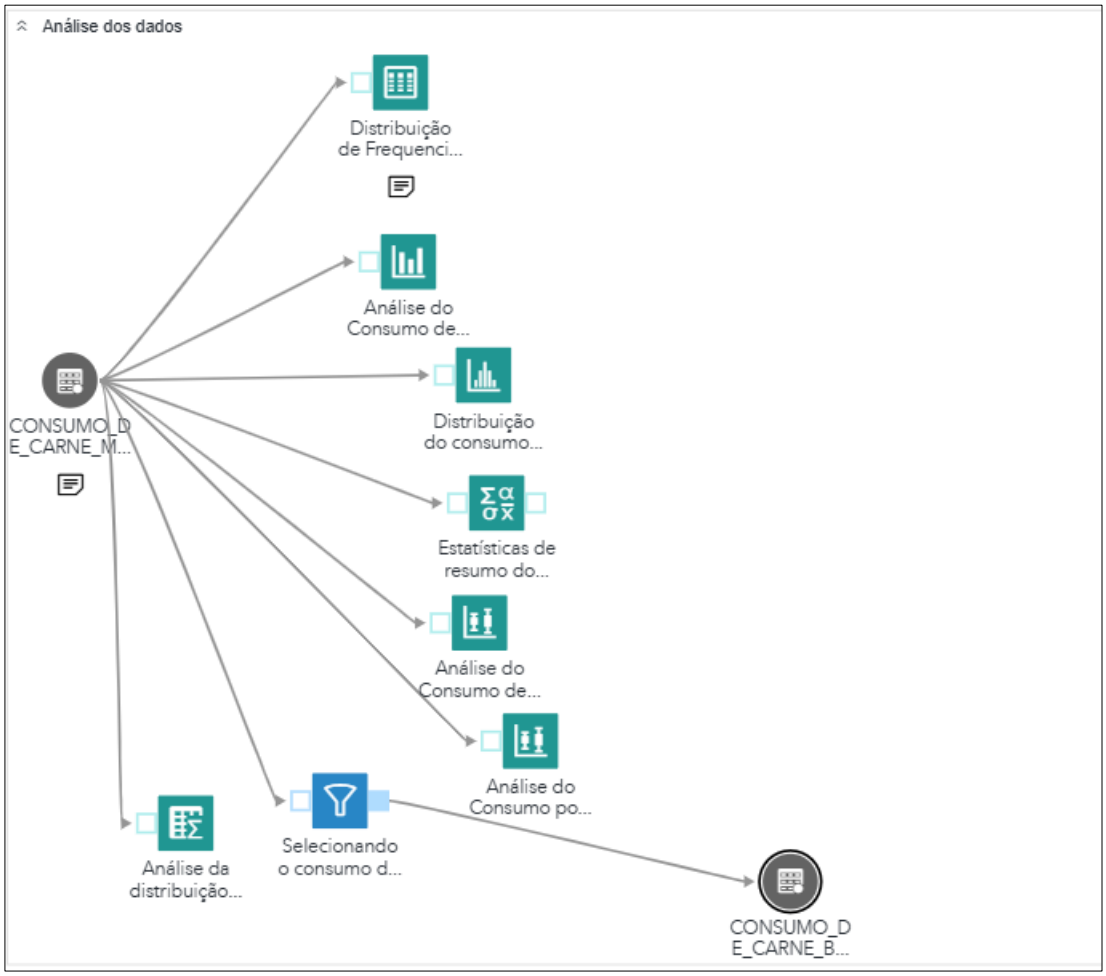
Para análise e exploração da base de dados de "***Consumo Mundial de Carne***" utilizaremos as soluções **SAS**[4] disponíveis no **SAS Viya 4 for Learners**[5] e desenvolveremos os seguintes ***Fluxos de Dados*** na solução ***SAS Data Studio (Develop Code and Flows).***

Figura 1: Fluxo de importação e preparação de dados para o formato SAS

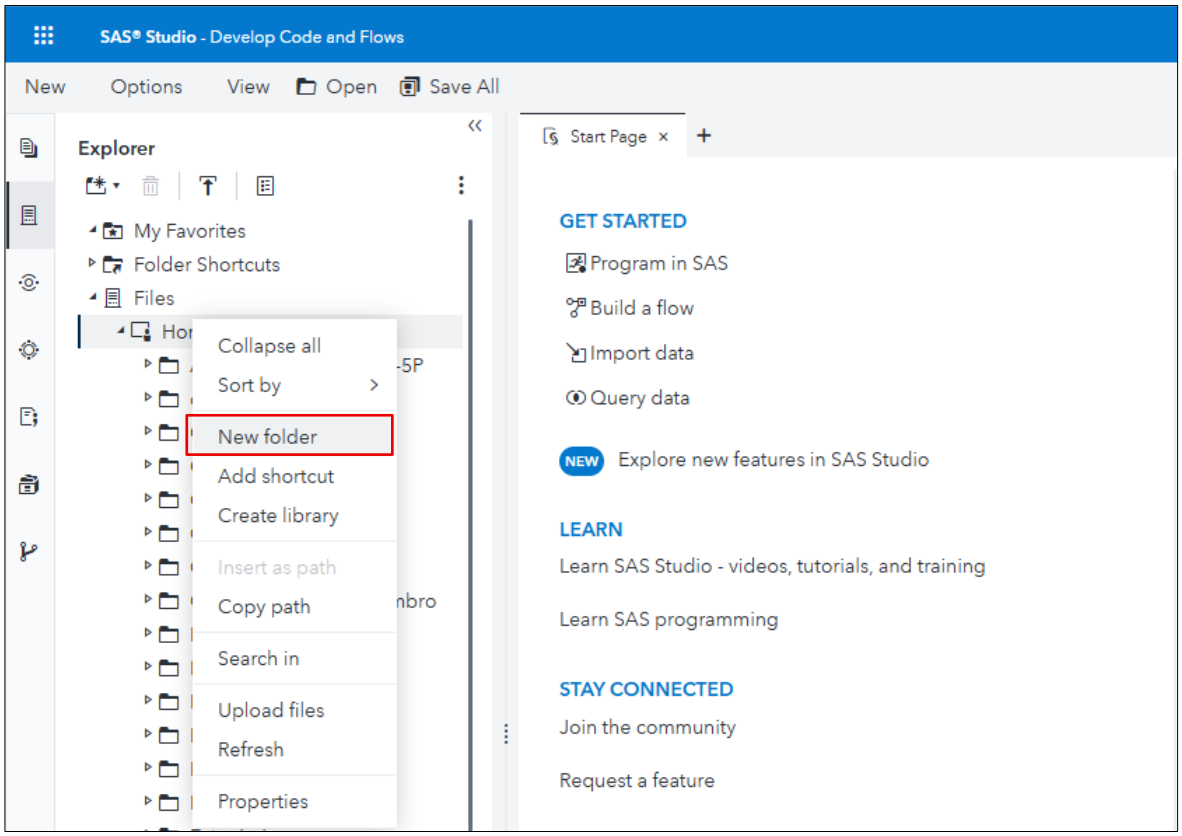


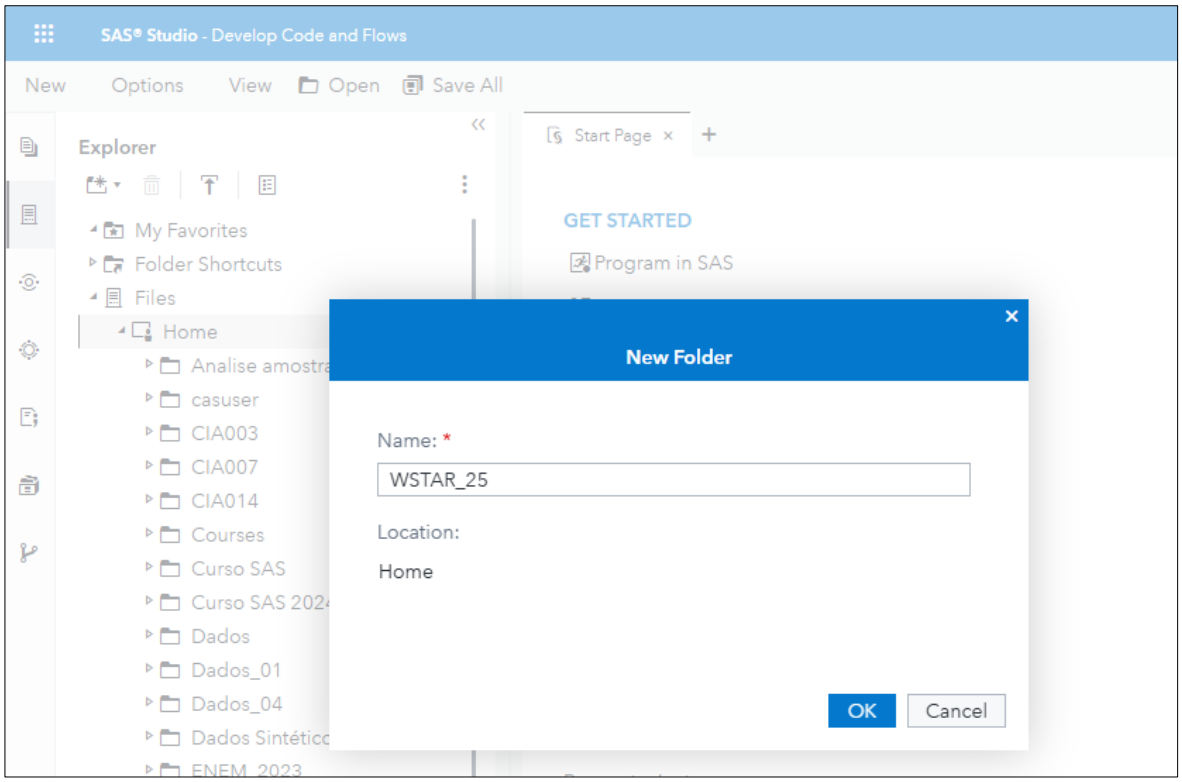
[4] https://www.sas.com/pt_br/home.html
[5] https://www.sas.com/en_us/software/viya-for-learners.html

Figura 2: Fluxo de exploração estatística dos dados

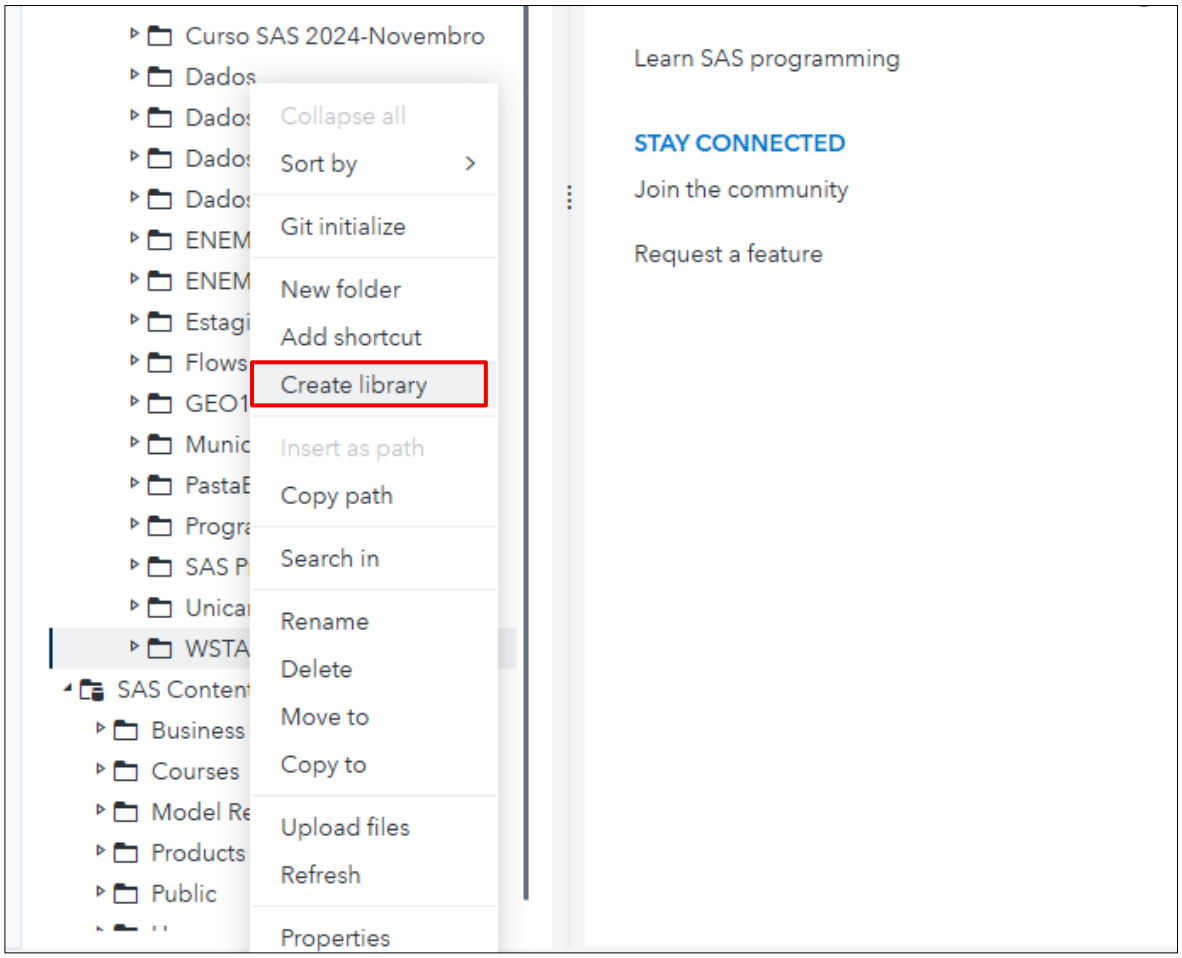


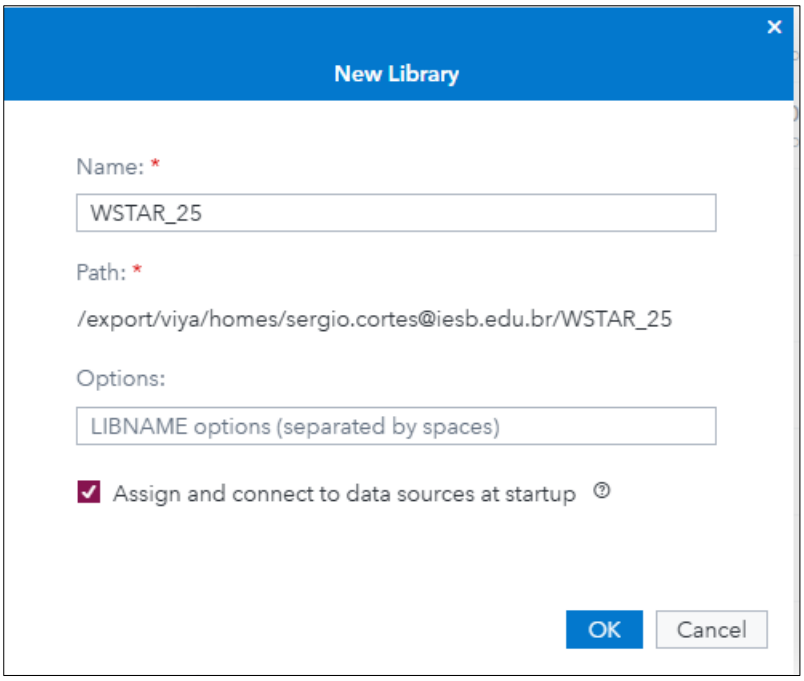
1. Criação da **pasta** para upload do arquivo de dados (clique com o botão direito do mouse na pasta “Home”.





2. Criação da **Library** para referência e utilização das bases de dados no Formato SAS (clique com o motão direito do mouse sobre a pasta criada no passo anterior)



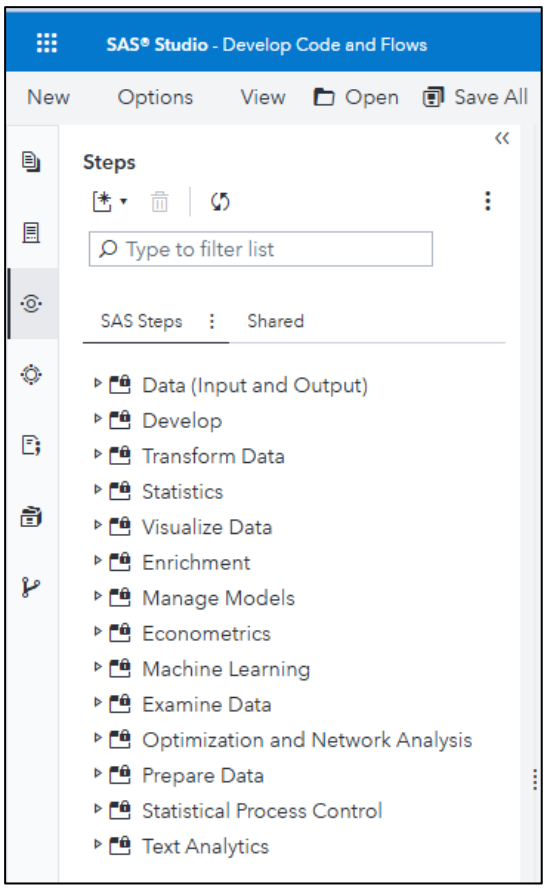


3. Steps do SAS Studio

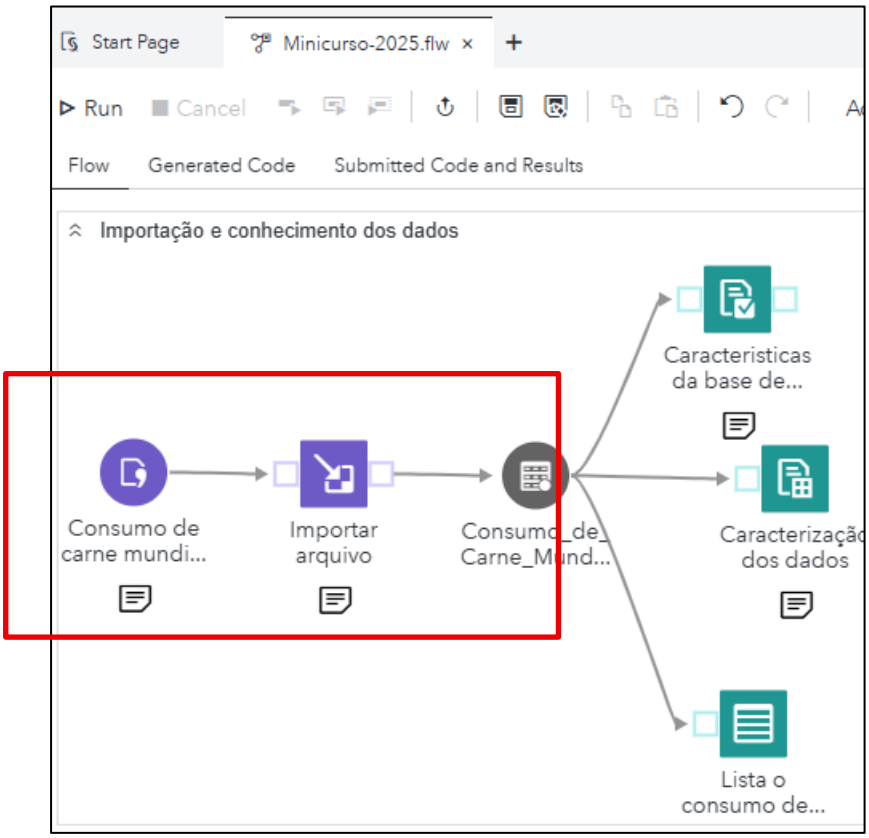
Os **Steps (etapas)** no **SAS Studio** são componentes gráficos que permitem ao usuário realizar tarefas de *análise de dados* por meio de uma interface intuitiva e sem a necessidade de escrever código. Cada **Step** representa uma atividade específica, como importar dados, calcular estatísticas, aplicar filtros, criar gráficos ou executar modelos. Eles são particularmente úteis para usuários iniciantes ou que desejam construir análises de forma visual, organizada e reprodutível.

Ao iniciar um projeto no **SAS Studio**, o primeiro passo do fluxo é geralmente a **importação dos dados**, utilizando o ***Step "Import Data"***, que permite carregar arquivos em formatos como CSV, Excel ou JSON a partir do seu diretório local ou servidor. Em seguida, o usuário pode adicionar o ***Step "Query Data"*** para selecionar variáveis de interesse, filtrar registros, criar colunas derivadas ou realizar junções com outras tabelas.

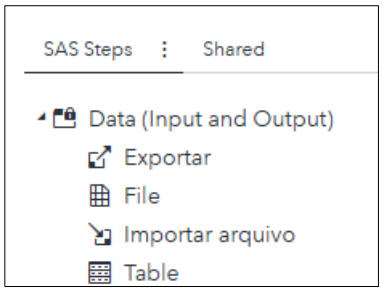
Selecione em cada grupo o **Step** adequado para execução do **NÓ** em seu **fluxo**.



4. Importação dos dados para o formato SAS

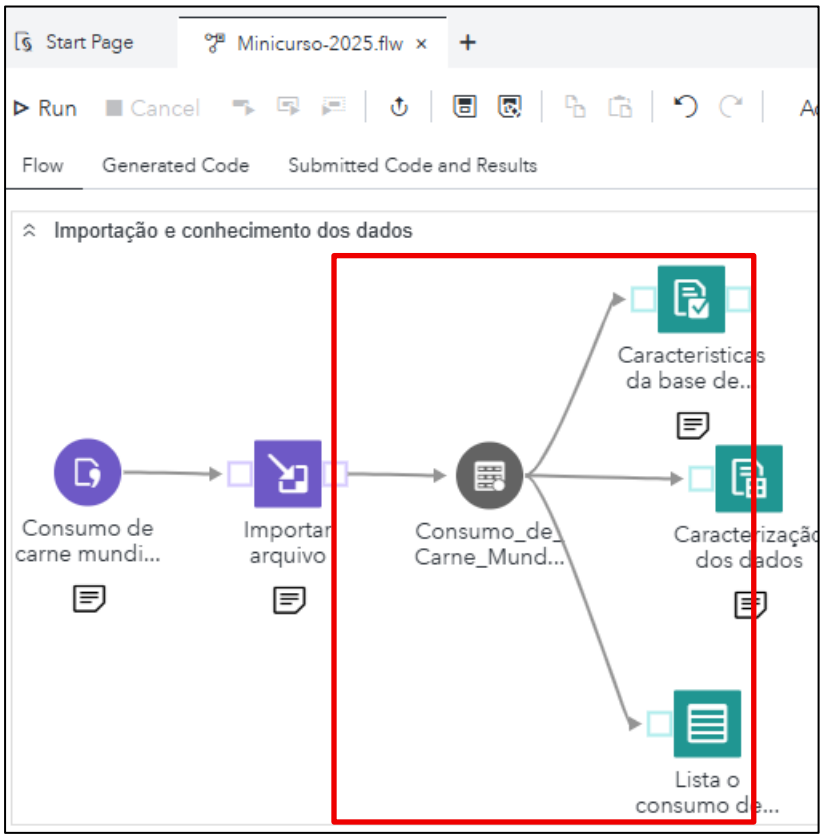


a. Importação dos dados para o formato SAS

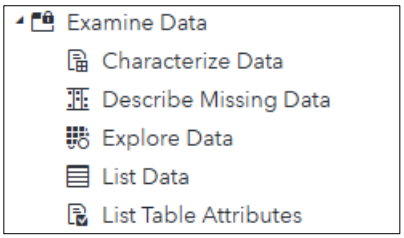


No grupo “Data (Input and Output) selecione “**File**”, “**Importar Arquivo**” e “**Table**”. Para cada Nó do seu fluxo defina os parâmetros conforme as especificações do arquivo que será importado.

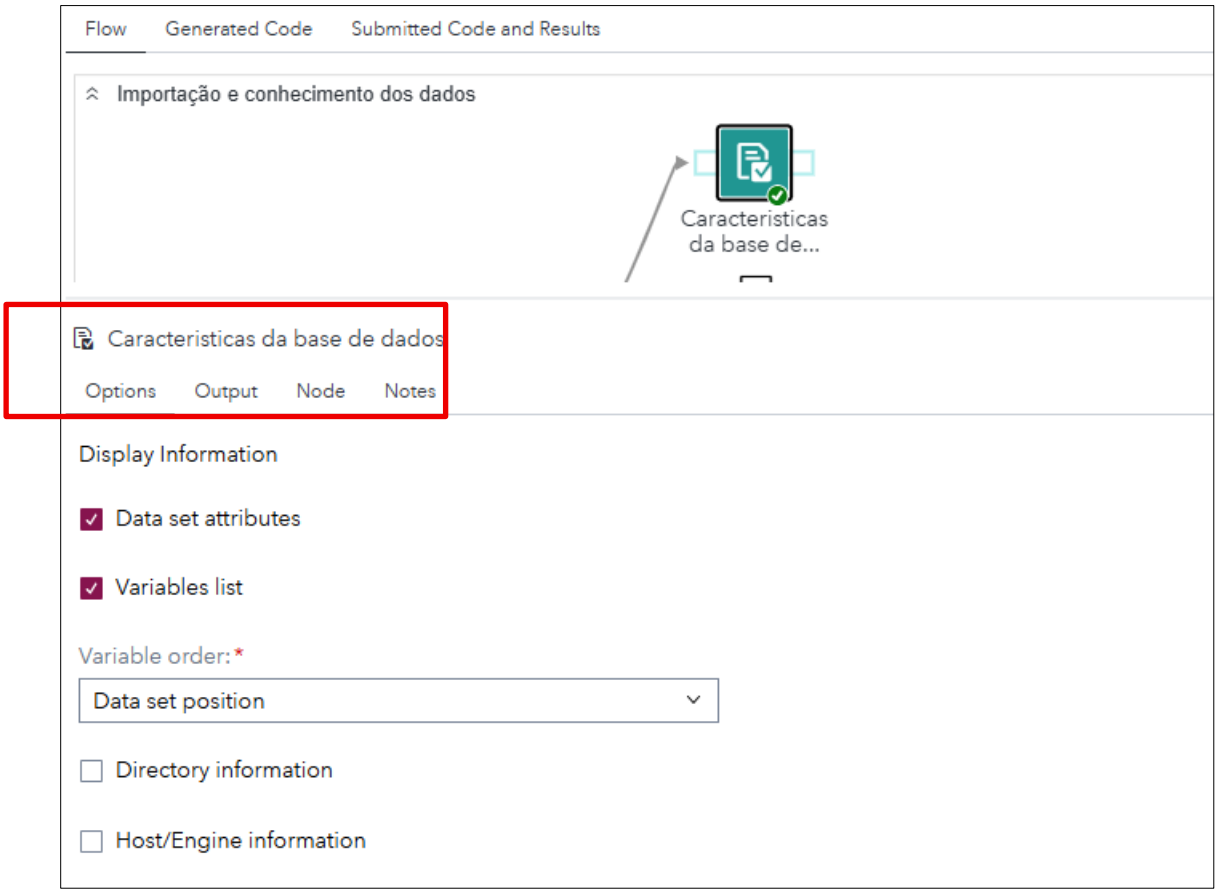
5. Exploração do arquivo convertido para o formato SAS



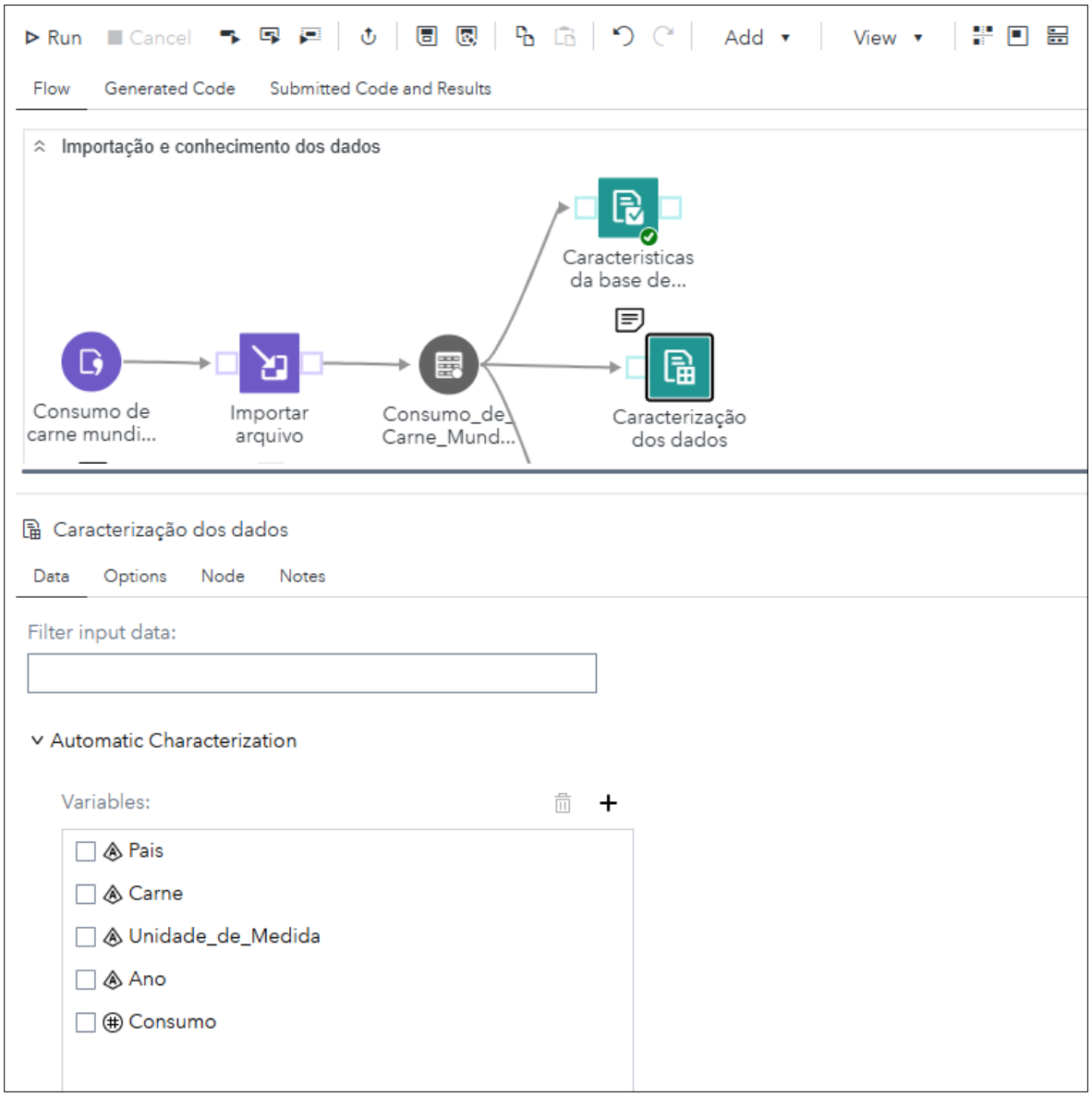
a. Utilize o grupo "**Examine Data**"



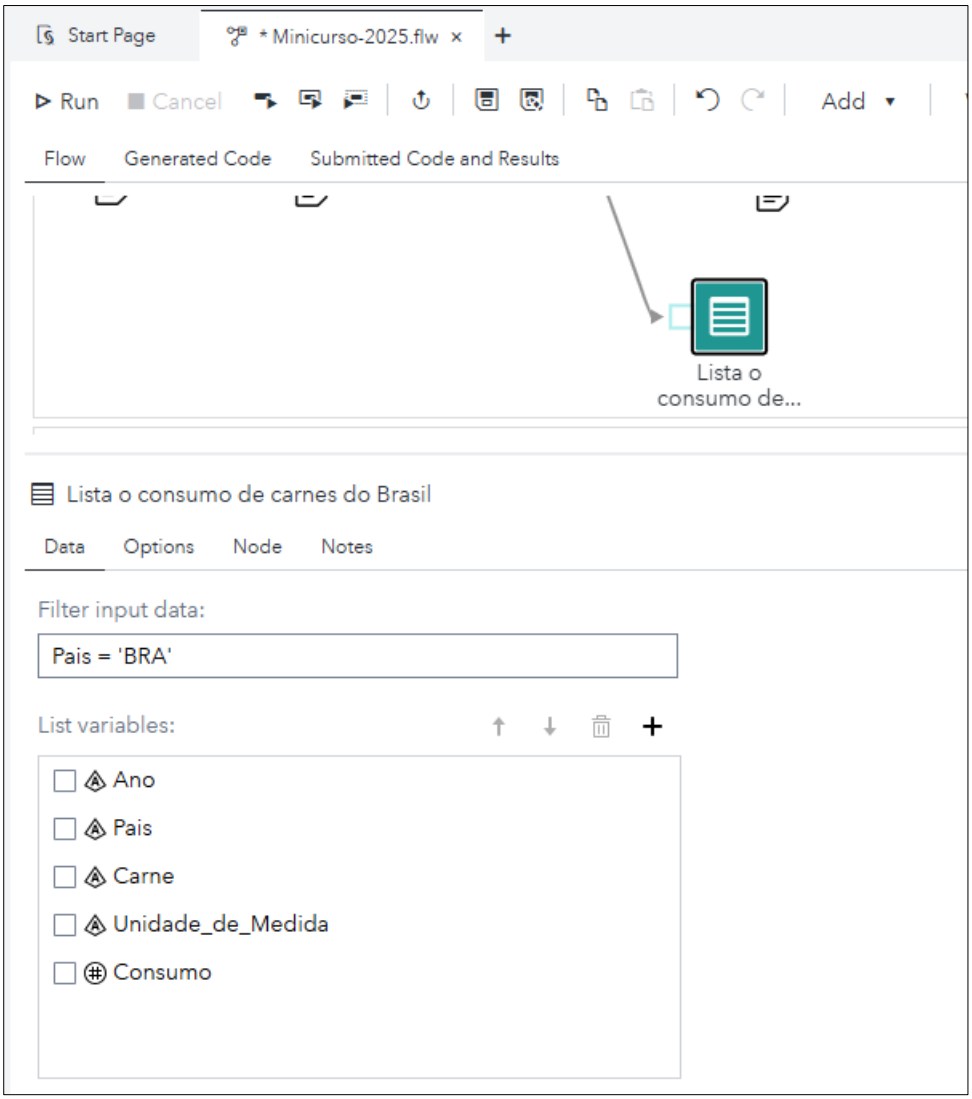
b. Utilize o **Step** "**List Table Attributes**" no seu fluxo e *avalie e preencha* de forma adequada as opções de cada aba.



c. Utilize o **Step** "**Characterize Data**" no seu fluxo para uma primeira análise dos dados importados.



d. Utilize o **Step** "**List Data"** no seu fluxo para selecionar um conjunto de dados do arquivo importado.



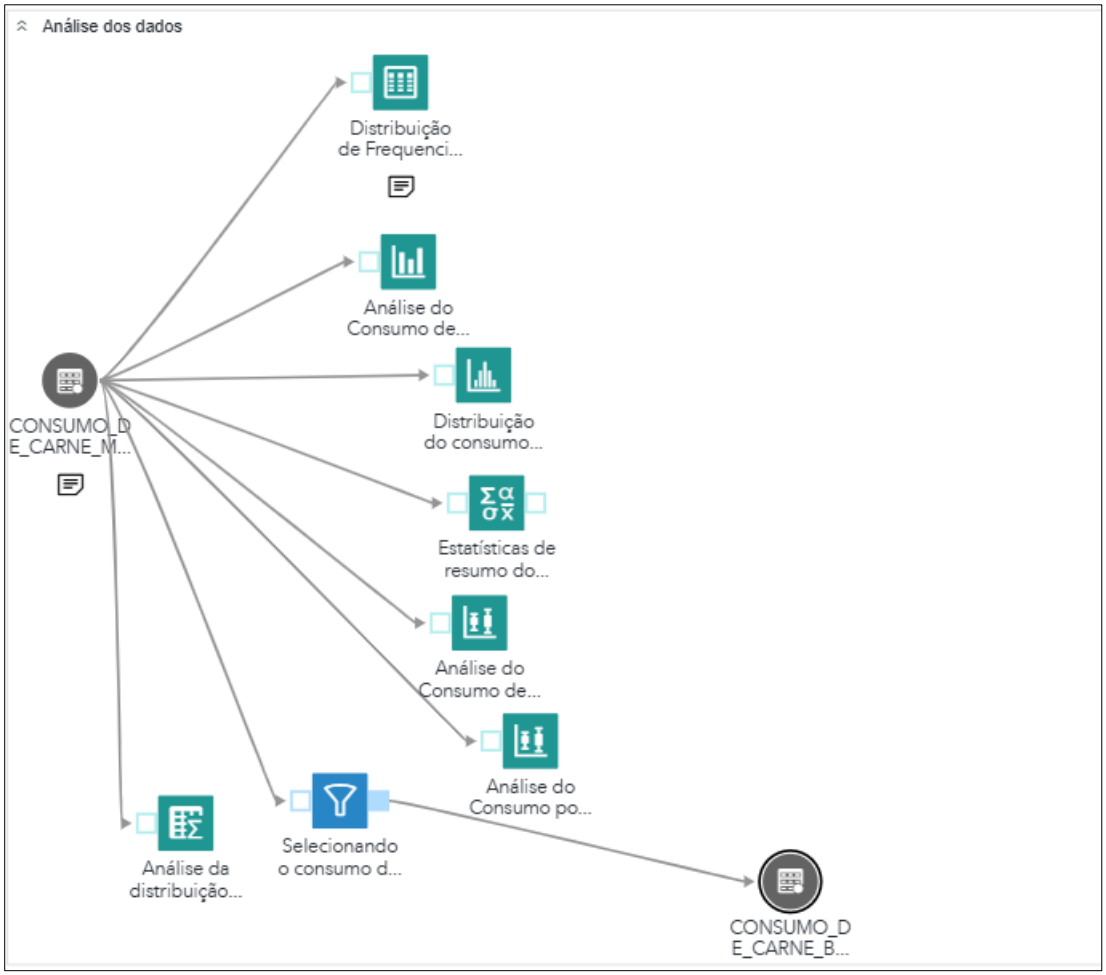
6. Análise da base de dados importada para o ambiente do SAS Studio

Depois da preparação dos dados, o próximo grupo de **Steps** pode incluir aqueles voltados à **análise estatística descritiva**, como:

- *Summary Statistics*: cálculo de médias, desvios padrão, mínimo, máximo etc.
- *Distribution Analysis*: histogramas e testes de normalidade.
- *Frequency Analysis*: tabelas de frequência para variáveis categóricas.

Além disso, o **SAS Studio** oferece **Steps** para **visualização de dados**, como gráficos de barras, linhas, dispersão, e até mapas geográficos, permitindo que os resultados sejam apresentados de forma clara e interativa. A seleção do tipo de **Step** depende do objetivo da análise e do tipo de variável em estudo.

Por fim, os **Steps** são organizados em **fluxos de trabalho encadeados** que facilitam a reprodutibilidade e o entendimento do processo analítico. O usuário pode salvar seu fluxo como um *Job* ou *Project*, reabrir posteriormente, executar novamente com novos dados, ou exportar os resultados. Isso torna o **SAS Studio** uma ferramenta poderosa e acessível para análises completas — *desde a entrada até a apresentação dos dados* — especialmente em aplicações como **Ciência e Tecnologia de Alimentos**, onde a interpretação estatística é essencial para decisões técnicas e científicas.



O grupo "**Statistics**" no **SAS Studio** organiza uma série de **Steps (etapas visuais)** projetadas para realizar análises estatísticas básicas a intermediárias sem a necessidade de digitar código. Esses Steps utilizam, em segundo plano, as tradicionais **procedures SAS (PROCs)** e são ideais para tarefas como frequências, estatísticas descritivas, testes de hipóteses, entre outros.

Entre os nós mais usados estão:

- **One-Way Frequencies** (PROC FREQ)
- **Summary Statistics** (PROC MEANS ou PROC SUMMARY)
- **Distribution Analysis** (PROC UNIVARIATE)
- **T Tests**, **ANOVA**, **Correlation**, **Regression**, etc.

Vamos destacar agora os dois nós mais utilizados no início da análise de dados:

I. One-Way Frequencies (Frequências Univariadas)

Este nó é utilizado para gerar **tabelas de frequência** de uma ou mais variáveis categóricas (qualitativas), apresentando a contagem de cada categoria, sua proporção relativa (percentual), frequência acumulada e outras medidas.

- **Procedure usada**: PROC FREQ
- **Aplicações comuns**:
  - ✓ Ver quantas observações existem para cada grupo de uma variável.
  - ✓ Identificar valores ausentes (missing).
  - ✓ Explorar a distribuição de variáveis qualitativas como sexo, cor, tipo de produto, origem, entre outros.

Exemplo prático em Ciência de Alimentos: Analisar a frequência de informações selecionadas por carne consumida.

II. Summary Statistics (Estatísticas de resumo)

Este nó calcula medidas descritivas para variáveis numéricas, como:

- Média
- Desvio padrão
- Mínimo e máximo
- Quartis
- Contagem de observações
- Procedure usada: PROC MEANS ou PROC SUMMARY
- Aplicações comuns:
  - ✓ Obter uma visão geral do comportamento das variáveis quantitativas.
  - ✓ Avaliar a variabilidade dos dados e identificar possíveis outliers.

- ✓ Comparar medidas centrais entre grupos (combinando com a variável de classificação).

Exemplo prático em Ciência de Alimentos: Calcular a distribuição da variável consumo de carnes em diversos países ao longo dos anos

III. Considerações Finais

Ambos os nós apresentados anteriormente são fundamentais no início de qualquer projeto de análise de dados, pois ajudam a **entender a estrutura dos dados, identificar inconsistências e orientar decisões subsequentes**. Como parte de um fluxo visual no **SAS Studio**, eles são facilmente combináveis com outras etapas como *Import Data*, *Filter Rows*, *Sort Data* e *Create Graph*, criando pipelines reprodutíveis e pedagógicos.

6 CONCLUSÃO

O uso das soluções SAS, em especial o **SAS Studio**, representa um avanço significativo na formação de profissionais e pesquisadores da área de **Ciência e Tecnologia de Alimentos**. Ao oferecer uma plataforma acessível, intuitiva e poderosa para análise estatística, visualização de dados e modelagem preditiva, o **SAS** contribui para o desenvolvimento de competências essenciais na interpretação de dados experimentais, controle de qualidade, inovação de processos e formulação de alimentos.

Ao longo deste minicurso, ficou evidente como o **SAS Studio** — com seus **Steps visuais**, **ferramentas de importação**, **estatísticas descritivas**, **gráficos automatizados** e **procedimentos avançados** — pode facilitar a análise rigorosa de dados laboratoriais, sensoriais, físico-químicos e microbiológicos, contribuindo diretamente para decisões baseadas em evidências.

O domínio dessas ferramentas fortalece a capacidade dos estudantes e profissionais de atuarem de forma mais crítica, analítica e inovadora, alinhando-se às boas práticas de pesquisa, desenvolvimento tecnológico e gestão da produção de alimentos. Mais do que aprender a usar um software, trata-se de integrar a

cultura analítica e a inteligência computacional ao cotidiano da **Ciência de Alimentos**, com o apoio de uma das plataformas mais consolidadas e confiáveis do mundo: o **SAS**.

Capítulo III

PLATAFORMA VIYA 4: ARQUITETURA MODERNA, KUBERNETES E SOLUÇÕES ANALÍTICAS EM ESCALA

Benjamin Farah[1]

DOI: 10.46898/home.9786560893306.3

1 Sr. Technical Architect, SAS Institute Brasil LTDA., Brasília , DF, email: benjamim.farah@sas.com. (SAS Brasil)

Resumo: A Plataforma Viya é uma solução *cloud-native* de *analytics* e IA que unifica acesso a dados, preparação, modelagem e visualização, oferecendo interfaces code-first e visuais para diferentes perfis de usuários; seu objetivo central é acelerar a geração de insights e decisões baseadas em dados. A arquitetura moderna da Viya é modular e orientada a serviços com APIs abertas, projetada para rodar em nuvem pública, privada ou híbrida; ao empacotar componentes em containers e orquestrá-los com Kubernetes, garante portabilidade, replicabilidade, resiliência e escalabilidade automatizada. As capacidades analíticas visam suportar processamento distribuído em larga escala, desde preparação de dados até machine learning e deep learning, com foco na operacionalização de modelos em produção, governança, auditabilidade e otimização de custo; a adoção exige planejamento por etapas para inventariar ativos, validar compatibilidade e mitigar riscos durante a migração.

Palavras-chave: Viya. Kubernetes. *Analytics*. Nuvem. *Machine Learning.*

Abstract: The Viya Platform is a cloud-native analytics and AI solution that unifies data access, preparation, modeling, and visualization, offering code-first and visual interfaces for different user profiles; its central purpose is to accelerate the generation of insights and data-driven decisions. Viya's modern architecture is modular and service-oriented with open APIs, designed to run on public, private, or hybrid clouds; by packaging components in containers and orchestrating them with Kubernetes, it ensures portability, reproducibility, resilience, and automated scalability. The analytics capabilities are intended to support large-scale distributed processing, from data preparation to machine learning and deep learning, with a focus on operationalizing models in production, governance, auditability, and cost optimization; adoption requires phased planning to inventory assets, validate compatibility, and mitigate risks during migration.

Keywords: *Viya. Kubernetes. Analytics. Cloud. Machine Learning.*

1 INTRODUÇÃO

A Plataforma SAS Viya posiciona-se no cerne das tecnologias de inteligência artificial como uma solução cloud-native que unifica dados, preparação, modelagem e visualização para acelerar a tomada de decisão baseada em dados. Ao oferecer interfaces **code-first** (Python, R) e ambientes visuais, Viya conecta cientistas de dados, engenheiros e decisores, permitindo que modelos e pipelines analíticos sejam desenvolvidos, validados e entregues com maior velocidade e consistência.

A sua arquitetura moderna baseada em microserviços e containers orquestrados por Kubernetes fornece portabilidade entre nuvens, escalabilidade horizontal, recuperação automática e isolamento de workloads, elementos essenciais para executar workloads de IA em produção com governança e auditabilidade. Essa fundação técnica reduz atritos na integração com código open source e infraestruturas de armazenamento escalável, facilitando a operacionalização de modelos, o monitoramento de performance e a gestão de custos.

Nas indústrias altamente reguladas e orientadas a anti-fraude, conformidade e riscos como o mercado financeiro, Viya entrega capacidades para detecção de fraude, gestão de risco, validação de modelos e conformidade; no varejo, permite personalização em tempo real, previsão de demanda e otimização de estoque; na mineração, potencializa análise preditiva para manutenção de ativos, otimização de cadeia logística e segurança operacional; já em telecomunicações, sustenta análise de churn, roteamento de rede e orquestração de serviços 5G. Em todos os setores, Viya atua como plataforma que traduz os avanços em IA em soluções práticas, governadas e escaláveis para desafios de eficiência, risco e experiência do cliente.

2 ARQUITETURA VIYA 4 E KUBERNETES

A arquitetura de Viya 4 é baseada em microserviços empacotados em containers e geridos por um operador de implantação Kubernetes, com recursos de infraestrutura como código para provisionamento do cluster e componentes de plataforma. O modelo de deployment usa charts Helm, Custom Resource Definitions e um SAS Viya Platform Deployment Operator para materializar configurações, disponibilizar serviços e aplicar updates de forma declarativa.

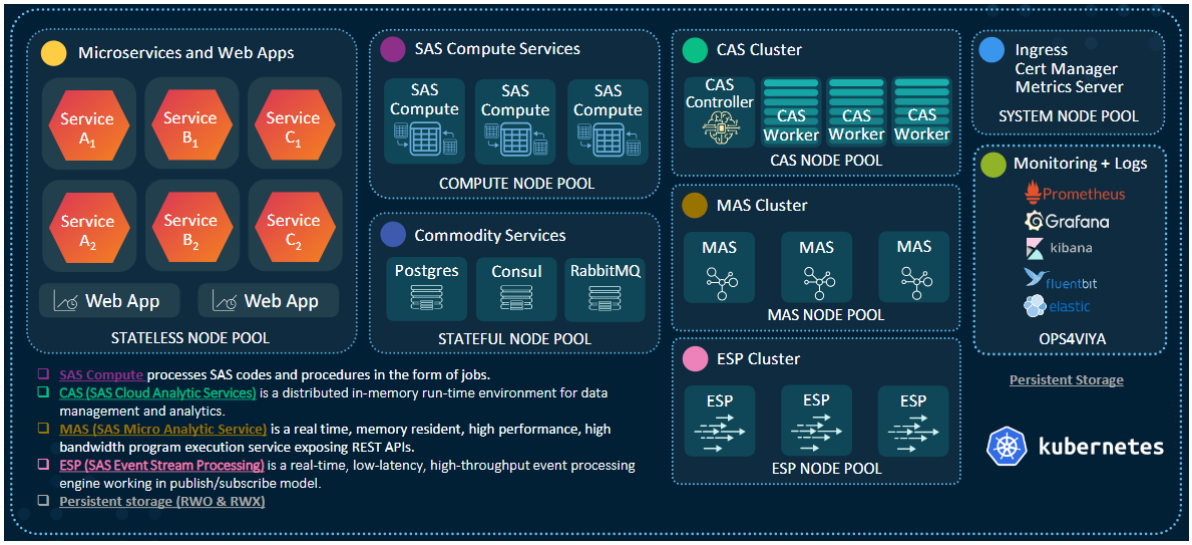


Figura 1 – Arquitetura da Plataforma SAS Viya 4.

3 VISÃO GERAL DOS MOTORES ANALÍTICOS DO SAS VIYA 4

A Plataforma SAS Viya 4 articula uma família de motores analíticos como serviços desacoplados e containerizados que executam funções complementares no ciclo de vida de dados e modelos. Cada motor tem responsabilidades claras: ingestão e preparação de dados, processamento analítico distribuído, treinamento e validação de modelos, inferência em lote e em tempo real, e governança de artefatos. A integração entre motores se dá por APIs REST, bytestreams de dados via CAS e repositórios de metadados versionados para garantir rastreabilidade e reprodutibilidade.

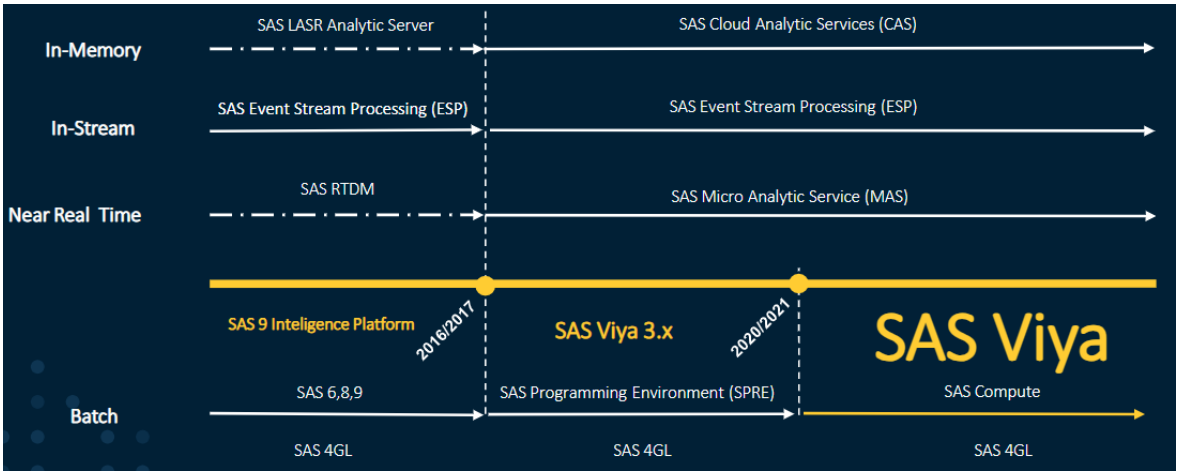


Figura 2 – Evolução dos Motores Analíticos da Plataforma SAS.

3.1 Principais motores, característica in-memory e aplicação recomendada

Tabela 1 – Motores Analíticos.

Motor	É in-memory?	Descrição curta	Aplicação mais adequada
SAS Cloud Analytic Services (CAS)	Sim	Motor distribuído in-memory projetado para processamento analítico paralelo e armazenamento temporário de tabelas e resultados.	Treinamento de modelos em larga escala; exploração interativa; visual analytics; pipelines de preparação e transformação de grandes volumes.
Batch Compute	Não	Ambiente tradicional de execução de código em lote, baseado em containers com acesso	Processamento batch, jobs agendados, pipelines ETL pesados e integração com legacy

Motor	É in-memory?	Descrição curta	Aplicação mais adequada
		a volume persistente e recursos de nós do cluster.	SAS code que não exige execução in-memory.
Event Stream Processing (ESP)	Não estritamente in-memory distribuído como o CAS; usa buffers e estados em memória local	Motor para análise contínua de fluxos, detecção de padrões complexos em tempo real e execução de regras em cima de eventos.	Monitoramento de fraude em tempo real, detecção de anomalias em telemetria e casos de resposta imediata a eventos.
SCR (SAS Container Runtime)	Não (conteúdo da imagem com runtime mínimo; execução local em container)	Runtime OCI para empacotar modelos/decisões em imagens imutáveis e portáveis; permite deploy independente do Viya em qualquer ambiente compatível com OCI.	Implantação e escalonamento de modelos como containers; validação de publicação; execução de scoring isolado e orquestrado por Kubernetes.
MAS	Depende da implementação (normalmente usa CAS para operações in-memory)	Serviço voltado para avaliação, scoring em massa e métricas de performance de modelos; integra pipelines de validação e comparações de runs.	Runs de avaliação em lote, geração de relatórios de performance e testes de regressão de modelos antes da promoção.

3.2 Como escolher o motor certo por requisito técnico

- Quando o objetivo primário for latência baixa e decisões em tempo real, selecione os componentes de inferência e decisioning containerizados integrados ao pipeline de eventos.
- Quando o requisito for exploração interativa, visual analytics e treinamento paralelo de modelos sobre grandes volumes, escolha o CAS como camada operacional principal e dimensione pools de memória e nós com CPU/GPU conforme necessidade.
- Para experimentação massiva, comparação de modelos e AutoML, utilize o runtime VDMML apoiado pelo CAS para paralelismo e velocidade.
- Para workloads batch e compatibilidade com código legado SAS, mantenha jobs e servidores de compute para execução orquestrada por Kubernetes CronJobs ou pipelines CI/CD.
- Para streaming e detecção de padrão contínua, utilize ESP com integração a mensageria e gates de decisão que acionam scorers ou workflows em CAS.
- Para governança, observabilidade e promoção, trabalhe com Model Manager e repositórios de metadados integrados ao pipeline CI/CD e ao operador de deployment do Viya.

3.3 Recomendações práticas de uso e integração entre motores

- Use o CAS como a camada de execução para exploração interativa e treino paralelo; direcione cargas de avaliação massiva do MAS para CAS pools dimensionados para memória e I/O; mantenha relatórios de avaliação no Model Manager para rastreabilidade.
- Empacote modelos aprovados com SCR para implantações imutáveis que podem ser orquestradas por Kubernetes, habilitando estratégias Canary, Blue/Green e escalonamento por réplicas de pod.
- Para scoring em tempo real com requisitos de latência rígidos, prefira APIs de scoring otimizadas e pequenas imagens SCR ou serviços de scoring

especializados; para batch scoring em massa, execute jobs controlados pelo MAS usando nós de compute ou CRONJobs Kubernetes.

- Integre MAS e SCR ao fluxo CI/CD: pipelines constroem e testam artefatos, MAS executa validações e relatórios automáticos, e SCR publica imagens que são promovidas por GitOps/ArgoCD para ambientes de validação e produção.

3.4 Impacto operacional e recomendações de infraestrutura

Recomenda-se que o cluster Kubernetes seja separado em pools de nós por função: nós otimizados para memória e I/O para CAS, nós com GPU para treino acelerado, nós de baixa latência para scorers e nós de armazenamento para batch persistence. Adoção de autoscaling horizontal para réplicas de serviços stateless e autoscaling de nós ou node pools para cargas stateful garante custo eficiente e desempenho previsível. Instrumente métricas de uso de memória CAS, latência de scoring e taxa de eventos ESP para políticas de SLO e automação CI/CD que incluam validações de desempenho antes de promoção para produção. Nós otimizados para memória e alta I/O cargas MAS de avaliação; nós com CPU/GPU para treino intensivo; pools leves e stateless para SCR e serviços de scoring em tempo real. Instrumente métricas específicas: consumo de memória CAS, throughput de execuções MAS, latência de SCR, taxa de erros e tempos de inicialização de containers para definir SLOs e políticas de autoscaling.

4 CAS

O SAS Cloud Analytic Services (CAS) é o motor in-memory distribuído da plataforma Viya que fornece o ambiente de execução para ingestão, preparação, processamento e análise paralela de grandes volumes de dados; ele atua como o plano de dados central onde tabelas são carregadas em memória e disponibilizadas para múltiplos consumidores via APIs e sessões SAS.

O funcionamento do CAS baseia-se em uma topologia com papéis bem definidos: um controller que aceita conexões e orquestra trabalho, workers que executam o processamento distribuído e um backup controller opcional para tolerância a falhas; a comunicação entre controller e workers permite dividir operações analíticas em tarefas paralelas e agregar resultados eficientemente.

Na prática, CAS mantém dados em memória para acelerar operações analíticas, mas usa um cache em disco (CAS disk cache) como overflow e suporte a persistência temporária; tabelas CAS são gerenciadas por políticas de rebalancing que permitem ajustar dinamicamente a distribuição dos blocos de dados quando os workers mudam, reduzindo a necessidade de recarregar dados após alterações de topologia em ambientes MPP.

Dois modos de implantação afetam comportamento e uso: SMP (Symmetric Multi-Processing) é o modo single-node onde o controller e o processamento ocorrem no mesmo pod e é indicado quando os conjuntos de dados são relativamente pequenos ou o ambiente é menos distribuído; MPP (Massively Parallel Processing) distribui a carga entre múltiplos worker pods, permitindo carregamento paralelo de dados e maior throughput em workloads de grande escala, porém com overhead de comunicação que em alguns cenários pode tornar SMP preferível.

Aplicações típicas do CAS incluem treino paralelo de modelos, exploração interativa e visual analytics, transformações em larga escala e pipelines de preparação de dados onde o acesso in-memory reduz latência e acelera iterações; em arquiteturas Viya ele costuma ser a camada central usada por runtimes como VDMML e por serviços de avaliação (MAS) para delegar computação intensiva e por scorers em batch quando se requer desempenho e throughput elevados.

Recomendações operacionais incluem provisionar nós dedicados e QoS garantido para pods CAS, dimensionar pools de memória e CAS disk cache conforme volumes esperados, considerar MPP quando houver grandes tabelas e múltiplos usuários concorrentes e habilitar políticas de table balancing para flexibilizar alterações no número de workers sem interrupções significativas de carga de trabalho.

5 ORQUESTRAÇÃO DE CONTAINERS E GERENCIAMENTO DE WORKLOADS

O cluster Kubernetes provê o controle de ciclo de vida dos pods Viya, scheduling baseado em labels e taints, tolerations para isolar workloads críticos e Node Affinity para mapear serviços a tipos de nó especializados. O operador Viya traduz recursos de alto nível em objetos Kubernetes que permitem rollback, reconciliação contínua e gestão de configurações sensíveis via Secrets e ConfigMaps. Ferramentas de CNI, storage class e provisionadores dinâmicos são aproveitadas para atender requisitos de rede e I/O dos componentes analíticos.

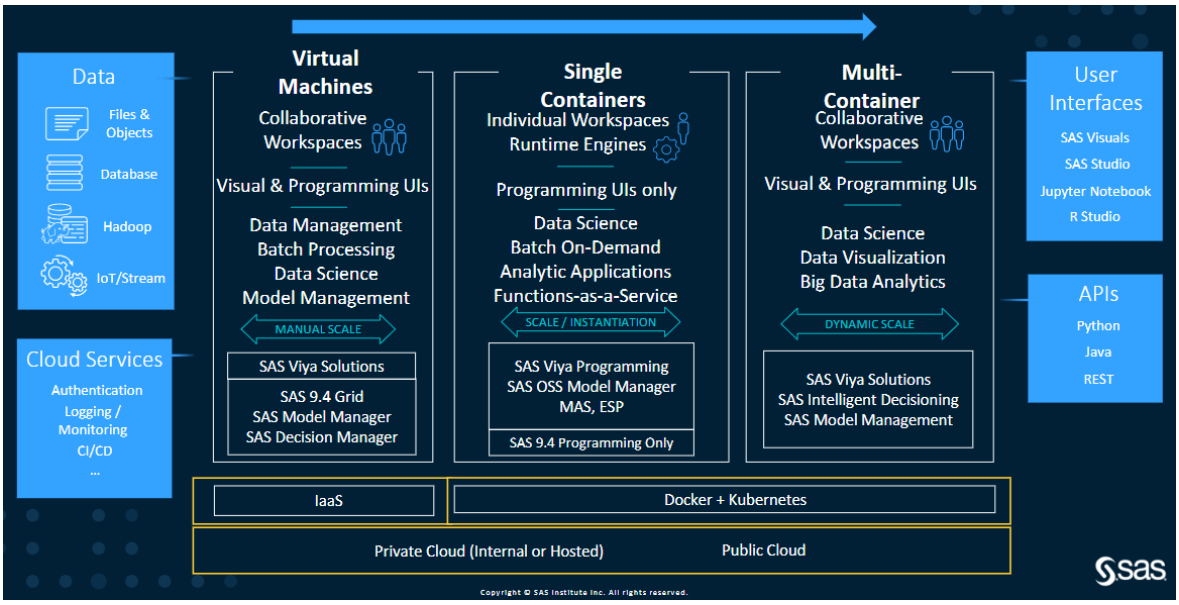


Figura 3 – Visão Geral da Plataforma SAS.

6 ESCALABILIDADE E GANHO DE DESEMPENHO EM TREINAMENTO, COMPARAÇÃO E ESCORAGEM DE MODELOS

A execução de cargas analíticas de Viya em Kubernetes habilita escalonamento horizontal automático de serviços stateless e escalonamento controlado de nós para workloads stateful, otimizando CPU, memória e GPUs conforme demanda. O SAS Cloud Analytic Services e motores distribuídos de Viya

beneficiam-se de múltiplos replicasets e recursos dedicados para reduzir latência de inferência, paralelizar treinamento e ampliar throughput de scoring. Workloads de treino podem ser direcionados a pools com aceleradores e nós com alta I/O para reduzir tempo de experimentação, enquanto pipelines de comparação e validação paralelizam testes A/B e runs concorrentes para acelerar seleção de modelos.

7 GOVERNANÇA, OBSERVABILIDADE E INTEGRAÇÃO CI CD

Viya 4 integra governança de modelos, audit trails e controle de acesso com APIs e serviços que persistem metadados e artefatos em repositórios centralizados. Observabilidade é entregue por métricas, logs e tracing que se integram a stacks de monitoramento Kubernetes para alerting e SLOs. A integração com CI CD usa imagens Docker customizadas, runners/executors Kubernetes e pipelines declarativos que automatizam build, testes unitários e de regressão, validação de modelos, promoção entre ambientes e deploys controlados via GitOps ou pipelines GitLab/ArgoCD. Políticas de aprovação, testes automatizados de conformidade e validação de performance são incorporadas ao fluxo para garantir entregas reprodutíveis e auditáveis.

8 CONCLUSÕES

Viya 4 sobre Kubernetes entrega uma plataforma unificada para desenvolvimento, comparação e operacionalização de modelos com escalabilidade sob demanda, isolamento de workloads e integração madura com práticas DevOps e governança corporativa. A combinação entre operadores, IaC e capacidades nativas de orquestração reduz o tempo de entrega de soluções analíticas em larga escala e fortalece controles necessários para ambientes regulados e críticos.

9 AGRADECIMENTOS

Agradeço ao SAS Brasil e em especial à Deborah Vasconcellos que como Sr. Global Academic Program Manager abriu a porta que permitiu a elaboração deste trabalho bem como a oportunidade de apresentar a tecnologia do SAS para o público da UFSM.

E a toda equipe organizadora do evento, em especial ao Renius Mello pela cordialidade e paciência na organização do workshop.

REFERÊNCIAS

[1] Sítio de Internet: documentation.sas.com, Acesso em 03/10/2025

[2] Sítio de Internet: communities.sas.com, Acesso em 03/10/2025

[3] Sítio de Internet: blogs.sas.com, Acesso em 03/10/2025

Capítulo IV

GERAÇÃO DE DADOS BIOLÓGICOS SINTÉTICOS USANDO MODELOS GENERATIVOS: UMA ABORDAGEM INOVADORA PARA ACELERAR NOVAS DESCOBERTAS

Tiago Bresolin[1]
Edgar Vargas Caballero[2]

DOI: 10.46898/home.9786560893306.4

1 Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, USA
2 Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, USA

Resumo: Modelos generativos tem se consolidado como uma ferramenta poderosa na ciência de dados, com destaque para as redes generativas adversariais (GAN), que aprendem a modelar distribuições complexas e produzir dados sintéticos realistas. No campo da genômica, essas redes têm despertado crescente interesse por sua capacidade de gerar dados *in silico* que preservam propriedades biológicas fundamentais, permitindo superar limitações relacionadas à escassez de dados, ao desbalanceamento de classes e à representação insuficiente de variantes raras. Além de ampliar conjuntos de dados genômicos, as GAN também oferecem novas possibilidades para simular cenários evolutivos, identificar assinaturas de seleção, otimizar predições genéticas e apoiar o desenvolvimento de estratégias de melhoramento mais robustas e orientadas por dados. Este capítulo apresenta os princípios fundamentais das GAN e exemplifica sua aplicação prática na geração de genótipos sintéticos de bovinos por meio de um modelo PC-WGAN com penalidade de gradiente. O modelo combinou análise de componentes principais com aprendizado adversarial para capturar a estrutura genética da população, reproduzindo padrões de desequilíbrio de ligação, distribuição de frequências alélicas e estrutura populacional observados nos dados reais. Os resultados demonstram que a abordagem é capaz de gerar amostras biologicamente plausíveis e estatisticamente consistentes, ampliando a diversidade genômica e oferecendo uma alternativa viável para estudos em populações de animais de produção. Em conjunto, os avanços revisados e o estudo de caso aqui apresentados destacam o potencial das GAN como ferramenta complementar em genômica pecuária, abrindo o caminho para uma integração mais ampla de modelos generativos na pesquisa e no melhoramento genético de precisão.

1 INTRODUÇÃO

Modelos generativos são um subconjunto dos algoritmos de aprendizado profundo (“deep learning”) desenvolvidos para produzir dados sintéticos que representam de forma realista as distribuições de dados reais (Hayawi et al., 2024). Esses modelos aprendem a distribuição subjacente dos dados observados, de modo a gerar novas amostras indistinguíveis das amostras reais (Rivero-Garcia et al.,

2024). Entre esses modelos, as redes generativas adversariais, também chamadas de “generative adversarial networks” (GAN), propostas por Goodfellow et al. (2014), tem se destacado como um dos algoritmos mais poderosas e amplamente utilizadas para a geração de dados sintéticos (Tripathi et al., 2022). Um GAN é composto por duas redes neurais em competição: o gerador, responsável por sintetizar dados a partir de amostras aleatórias, e o discriminador, encarregado de distinguir entre amostras reais e amostras geradas (Goodfellow et al., 2014). Por meio desse processo de treinamento adversarial, os GAN refinam iterativamente a qualidade dos dados produzidos até que eles se tornem praticamente indistinguíveis dos dados reais (Gangwal & Lavecchia, 2024). Essa estrutura robusta já foi aplicada com sucesso em diferentes áreas, incluindo processamento de linguagem natural (Miller et al., 2022; Rannon & Burstein, 2025), visão computacional (Olaniyi et al., 2022), síntese de vídeo (Aldausari et al., 2023) e geração de áudio (Haque et al., 2020; Liao et al., 2024).

Dado o êxito na geração de dados nas áreas mencionadas, os GAN também têm demonstrado potencial para gerar dados biológicos e moleculares, acelerando pesquisas nas áreas de farmacologia e genômica (Tripathi et al., 2022; Murad et al., 2023). Por exemplo, Tamilmani et al. (2022) desenvolveram um modelo para sintetizar sequências de microRNA como biomarcadores para detecção de câncer, mostrando que a combinação de dados reais e sintéticos pode aprimorar o desempenho de classificadores em análises de expressão gênica. Hazra et al. (2022) aplicaram um método baseado em GAN para gerar sequências sintéticas do genoma de gatos, obtendo correlação média de 93,7% com a sequência genômica real da espécie, evidenciando o potencial desses algoritmos para impulsionar pesquisas genômicas. Já Yelmen et al. (2021) mostraram que genomas artificiais gerados por GAN aumentaram em 4,4% a acurácia da imputação, especialmente para alelos raros, quando utilizados para expandir populações de referência. Esses avanços demonstram o poder dos GAN na geração de dados genômicos e abrem caminho para aplicações em animais de produção, onde dados genômicos sintéticos podem suprir lacunas críticas de disponibilidade e diversidade de dados.

Apesar dos progressos recentes no uso de GAN para gerar dados genômicos em diferentes espécies, sua aplicação para gerar dados genômica em animais de

produção não tem sido ainda reportado na literatura. Entre os principais desafios estão garantir a validade biológica dos dados gerados e capturar de forma fidedigna a complexa arquitetura genética e a diversidade intrínseca às populações de interesse (Szatkownik et al., 2023). Superar essas limitações poderá abrir novas fronteiras na genômica animal, permitindo a simulação de combinações genéticas diversas e apoiando o desenvolvimento de estratégias de seleção mais eficientes para características economicamente relevantes em animais de produção. Além disso, dados gerados por GAN poderão ser aproveitados para otimizar protocolos de edição gênica, ao prever e minimizar efeitos fora do alvo (off-target), aumentando a precisão da engenharia genômica. Nosso objetivo nesse capítulo é apresentar princípios e o uso de modelos generativos para gerar dados biologicamente plausíveis em genômica de animais de produção. Questões matemáticas ou detalhes técnicos de treinamento do GAN não serão discutidos aqui, pois já foram amplamente cobertos em trabalhos já disponíveis na literatura (Aggarwal et al., 2021; Nayak et al., 2024). Para leitores interessados em uma compreensão mais aprofundada da fundamentação técnica, recomenda-se a consulta dessas referências.

2 PRINCÍPIOS DAS REDES GENERATIVAS ADVERSARIAIS

Modelos generativos constituem um ramo do aprendizado profundo (“deep learning”) projetado para produzir dados sintéticos que reflitam as propriedades estatísticas de conjuntos de dados reais (Yelmen & Jay, 2023). Diferentemente dos modelos discriminativos, cujo foco está em aprender características que distinguem entre categorias (Nayak et al., 2021), os modelos generativos buscam aprender uma representação latente da distribuição subjacente dos dados, permitindo a geração de novas amostras estatisticamente indistinguíveis das reais (Shi et al., 2023). Diversas arquiteturas se enquadram na categoria de modelos generativos, incluindo autoencoders (Joo et al., 2020), máquinas de Boltzmann (GM et al., 2020) e os generative adversarial networks (GAN) (Goodfellow et al., 2014). Entre essas abordagens, o modelo GAN têm atraído considerável atenção devido à sua capacidade de modelar distribuições complexas e de alta

dimensionalidade, gerando dados altamente realistas. Esta seção apresenta os princípios fundamentais dos GAN, que servem de base conceitual para este capítulo, com ênfase especial em seu potencial para a geração de dados genômicos artificiais aplicados à animais de produção.

2.1 Arquitetura

Proposto por Goodfellow et al. (2014), os GAN baseiam-se no princípio do treinamento adversarial, no qual dois modelos competem em um jogo de soma zero. Essa técnica tem um paralelo conceitual com a coevolução biológica, em que espécies que interagem se adaptam continuamente em resposta às mudanças umas das outras para manter vantagem competitiva (Stewart et al. 2025; Papkou et al., 2019). Em sistemas hospedeiro-parasita, por exemplo, os parasitas podem evoluir seus mecanismos para escapar das defesas imunológicas do hospedeiro, o que leva o hospedeiro a desenvolver novas estratégias de defesa, estimulando, por sua vez, novas adaptações parasitárias (Carvalho et al., 2024). A ideia de competição adversarial em aprendizado de máquina teve início em 1992, quando Schmidhuber propôs sistemas compostos por redes em competição que se auto aperfeiçoavam (Schmidhuber, 1992). Mais tarde, Li et al. (2013) aplicaram princípios coevolutivos em um contexto de modelagem generativa para simular comportamentos em animais, treinando um modelo para gerar comportamentos sintéticos e outro para classificá-los. Essa competição e aprendizado iterativo permitiu que o classificador distinguisse progressivamente entre os comportamentos reais dos gerados, mostrando a dinâmica do aprendizado adversarial.

Esses avanços biológicos e computacionais culminaram na formalização dos GAN, em que duas redes neurais, a geradora e o discriminadora, são treinadas simultaneamente dentro de uma estrutura competitiva (Goodfellow et al., 2014). Conforme ilustrado na Figura 1, a rede geradora inicia amostrando aleatoriamente de um espaço latente e transforma essas amostras em dados sintéticos similares as observações reais. Esse espaço latente, uma representação de menor dimensão da distribuição de dados-alvo, é progressivamente otimizado durante o treinamento para produzir dados que se aproximam cada vez mais às propriedades estatísticas

dos dados reais (Alqahtani et al., 2021; Wolterink et al., 2021). Inicialmente, a rede geradora produz dados irreais e distante da distribuição dos dados reais. Esses dados gerados, junto a amostras reais, são fornecidos a rede discriminadora, um classificador binário que atribui alta probabilidade aos dados reais e baixa probabilidade aos dados sintéticos (Shafahi et al., 2019). A retroalimentação da rede da rede discriminadora, calculado por meio de uma função de perda, é propagado de volta por ambas as redes. Usando a técnica de "*backpropagation*", a rede geradora atualiza seus parâmetros para minimizar seus erros, aprimorando sua capacidade de produzir dados sintéticos que a rede discriminadora não consegue distinguir de exemplos reais (Goodfellow et al., 2014).

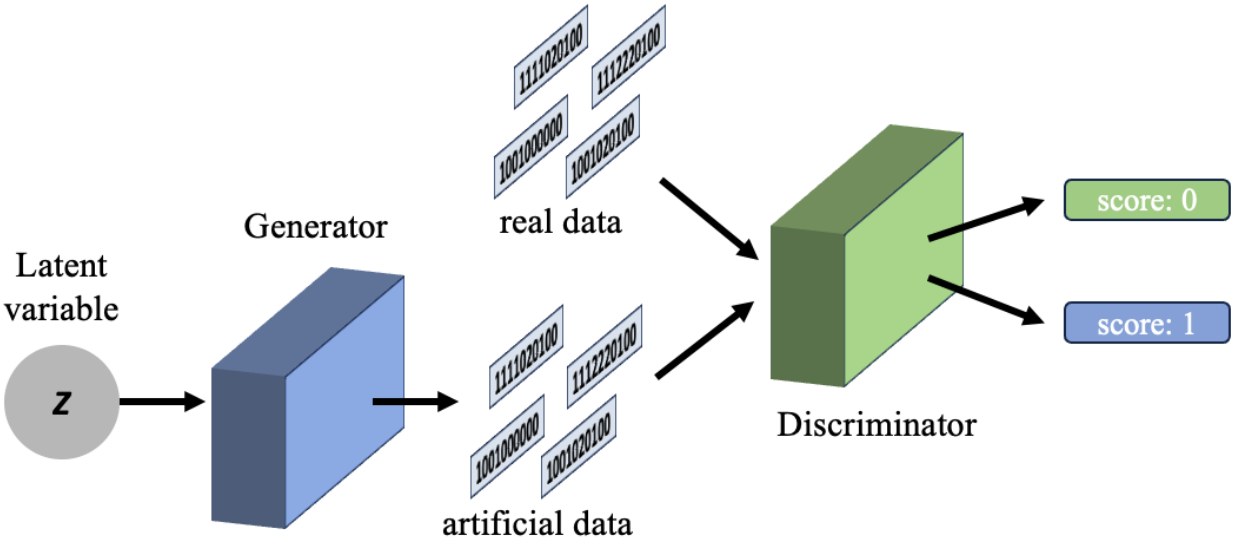


Figure 1. Visão geral da estrutura de um model de redes generativas adversariais para a geração de dados sintéticos.

O objetivo da rede discriminadora é classificar corretamente os dados produzidos pela rede geradora como dados reais ou sintéticos, enquanto o objetivo da rede geradora é produzir amostras sintéticas que a rede discriminadora classifique erroneamente como reais. Durante o treinamento, ambas as redes melhoram iterativamente por meio dessa interação adversarial. No ponto de convergência, a rede discriminadora passa a atribuir probabilidade de 0,5 a todas as amostras, indicando que já não consegue distinguir com confiança se os dados gerados são reais ou sintéticos. Esse estado, conhecido como equilíbrio de Nash, reflete o momento em que a rede geradora aprendeu efetivamente a distribuição dos dados e a rede discriminadora encontra-se em máxima incerteza (Thirumagal

& Saruladha, 2021; Wenzel, 2023). A arquitetura dos GAN é particularmente adequada para gerar dados moleculares e biológicos de alta dimensionalidade e estrutura complexa. No contexto da genômica de animais de produção, GAN têm o potencial de gerar sequências genômicas sintéticas que preservam características genéticas essenciais, como padrões de desequilíbrio de ligação e diversidade de haplótipos. Esses dados podem auxiliar em aplicações como a detecção de assinaturas de seleção, a otimização de estratégias de melhoramento e o refinamento de protocolos de edição gênica.

2.2 Função de perda

As funções de perda desempenham um papel central em garantir que os GAN gerem amostras sintéticas confiáveis e biologicamente relevantes (Yıldız et al., 2024). No treinamento de GAN, a rede geradora e a discriminadora são otimizadas dentro de uma estratégia competitiva que impulsiona cada modelo em direção à sua máxima eficiência: a rede geradora busca criar dados sintéticos indistinguíveis das observações reais, enquanto a rede discriminadora procura identificar corretamente a verdadeira origem dos dados (Yıldız et al., 2024; Dong & Yang, 2019). Esse processo adversarial é guiado pela função de perda, que quantifica a distância entre a distribuição das amostras geradas e a dos dados reais (Pan et al., 2020; Trevisan de Souza et al., 2023). Na prática, o objetivo geral é decomposto em duas funções de perda: uma para a rede geradora e outra para a discriminadora (Trevisan de Souza et al., 2023). Essa separação fornece gradientes úteis para o "*backpropagation*", permitindo que a rede geradora aproxime progressivamente a distribuição real dos dados, ao mesmo tempo em que orienta a rede discriminadora a refinar sua capacidade de classificar a origem das amostras (Goodfellow et al., 2014).

Goodfellow et al. (2014) propuseram como função de perda para os GAN a função binária de entropia cruzada ("*binary cross-entropy*"), que se baseia na divergência de Jensen-Shannon para medir a similaridade entre a distribuição de dados reais e sintéticos (Sinn & Rawat, 2017). No entanto, essa formulação frequentemente gera instabilidades no treinamento, já que a rede discriminadora pode se tornar muito poderoso em relação a rede geradora, dificultando sua

evolução (Saxena & Cao, 2020). Para mitigar essas limitações, diversas funções de perdas alternativas foram propostas com o objetivo de melhorar a estabilidade e a dinâmica durante o treinamento dos GAN (Dong & Yang, 2019). Por exemplo, Arjovsky et al. (2017) propuseram substituir a divergência de Jensen-Shannon como função de perda pela distância de Wasserstein, que mede o custo mínimo de transformar uma distribuição de probabilidade em outra (Saqlain et al., 2021; Okano & Imaizumi, 2022). Nessa abordagem, o discriminador é redefinido como um crítico, cuja função é estimar a distância de Wasserstein em vez de realizar classificação binária (Song & Ermon, 2020). A Wasserstein GAN (WGAN) fornece gradientes mais informativos para o gerador, mesmo quando há pouco ou nenhuma sobreposição entre as distribuições real e sintética, resultando em maior estabilidade durante o treinamento e melhor qualidade das amostras geradas (Saqlain et al., 2021).

Para aumentar ainda mais a estabilidade durante o treinamento dos GAN, Gulrajani et al. (2017) sugeriram a Wasserstein GAN com penalização de gradiente (WGAN-GP), que impõe a continuidade de Lipschitz (propriedade matemática) ao penalizar desvios na norma do gradiente do discriminador. Essa modificação estabiliza a otimização da rede discriminadora e reduz problemas de subdesempenho, embora o custo a demanda computacional seja maior (Yıldız et al., 2024; Lu, 2024). Apesar do tempo de treinamento mais longo, a WGAN-GP mostrou produzir amostras mais diversas e robustas em comparação à WGAN padrão (Lu, 2024). Outras variações de funções de perdas também foram propostas. Mao et al. (2017) desenvolveram a Least-Squares GAN (LSGAN), que substitui a entropia cruzada binária pelo erro quadrático médio, reduzindo a divergência qui-quadrado de Pearson. Essa modificação aumenta a estabilidade e reduz a sensibilidade ao ruído (Yıldız et al., 2024). De forma semelhante, Qi (2017) apresentou a Loss-Sensitive GAN (LS-GAN), que utiliza uma função de perda baseada em margens, garantindo que amostras reais tenham valores de perda menores que as sintéticas. Esse ajuste melhora a generalização e a síntese de novas amostras (Saqlain et al., 2021; Iglesias et al., 2024). Portanto, esses avanços destacam a importância de selecionar uma função de perda adequada à arquitetura e às características do conjunto de dados. A escolha da função de perda impacta diretamente a estabilidade dos GAN, a convergência do treinamento e a relevância

biológica das amostras geradas (Dong & Yang, 2019; Lucic et al., 2018; Pan et al., 2020).

2.3 Desafios

Embora os GAN tenham demonstrado notável potencial para gerar dados sintéticos realistas, trainar esses algoritmos continua sendo um dos maiores desafios para sua aplicação em larga escala. As dinâmicas adversariais entre a rede geradora e a discriminadora frequentemente resultam em instabilidade, desequilíbrio e dificuldades de convergência. Além disso, problemas como colapso e a ausência de métricas universais de avaliação da qualidade dos dados gerados dificultam ainda mais seu uso, especialmente em domínios como a genômica, nos quais tanto a fidelidade biológica quanto a diversidade em nível populacional são essenciais. Nesta seção, iremos revisar três desafios principais sendo eles: instabilidade e equilíbrio no treinamento, colapso, e avaliação da fidelidade dos dados gerados pelos GAN, destacando suas implicações para aplicações em genômica para animais de produção.

2.3.1 Instabilidade e equilíbrio no treinamento

Um dos objetivos centrais durante o treinamento dos GAN é alcançar o equilíbrio de Nash, definido como estado no qual o discriminador não consegue distinguir de forma confiável entre dados reais e sintéticos e passa a atribuir probabilidades iguais a ambos (Wolterink et al., 2020). Na prática, atingir esse equilíbrio é extremamente difícil porque a rede geradora e a discriminadora estão presos em uma competição dinâmica: enquanto a rede geradora busca criar dados capazes de enganar a rede discriminadora, a rede descriminadora trabalha para classificar corretamente dados reais versus sintéticos (Salimans et al., 2016). Essas forças opostas frequentemente levam a treinamentos instáveis dos GAN, nos quais nenhuma das redes converge para um estado ótimo, resultando em amostras irreais ou soluções subótimas (Khanuja & Khanuja, 2021).

Uma das principais fontes de instabilidade é o sobreajuste do discriminador. Quando a rede discriminadora classifica dados sintéticos como falsos com quase

100% de acurácia, os gradientes se tornam nulos ("*vanishing gradientes*"), privando a rede geradora de retroalimentação útil e dificultando sua evolução ou aprendizado (Goodfellow et al., 2014). Esse excesso de otimização também restringe a capacidade de generalização da rede discriminadora, reduzindo sua eficácia em guiar as atualizações da rede geradora (Wolterink et al., 2020). Para evitar esse problema, Goodfellow et al. (2014) recomendaram atualizações alternadas entre as redes, uso de taxas de aprendizado moderadas (tipicamente 0,0002) e múltiplas atualizações da rede discriminadora para cada atualização da rede geradora (Ferreira et al., 2024). O pré-treinamento da rede discriminadora apenas com dados reais também foi sugerido como forma de estabilizar o treinamento nas iterações iniciais (Hiriyannaiah et al., 2020). Por outro lado, se a rede geradora começa a superar a rede discriminador, o sinal de retroalimentação deste perde relevância, novamente desestabilizando o processo de treinamento (Ferreira et al., 2024).

Além das estratégias de treinamento mencionados anteriormente, avanços nas funções de perda foram fundamentais para melhorar a estabilidade dos algoritmos durante o treinamento. Goodfellow et al. (2014) propuseram modificar o objetivo da rede geradora para maximizar a probabilidade de ser classificado como real, evitando gradientes nulos. Um desenvolvimento ainda mais impactante foi a Wasserstein GAN (WGAN), que substituiu a divergência de Jensen-Shannon pela distância de Wasserstein (Arjovsky et al., 2017). Essa métrica fornece gradientes mais suaves e retroalimentação mais informativos para a rede geradora, mesmo quando há pouca sobreposição entre distribuições reais e sintéticas, melhorando as propriedades de convergência (Wolterink et al., 2020). Posteriormente, Gulrajani et al. (2017) introduziram a penalização de gradiente (WGAN-GP) para impor a continuidade de Lipschitz, prevenindo explosão ou desaparecimento de gradientes e trazendo maior estabilidade, embora com maior custo computacional (Yilmaz & Korn, 2024). Esses avanços metodológicos são especialmente importantes na genômica, onde os algoritmos precisam lidar com conjuntos de dados de alta dimensionalidade e estruturas populacionais complexas.

2.3.2 Colapso

Além do desafio da instabilidade durante o treinamento dos GAN, outro problema recorrente é o colapso de modelos ("*mode collapse*"), situação em que a rede geradora passa a produzir dados muito semelhantes entre si, sem capturar toda a diversidade da distribuição de dados reais (Ferreira et al., 2024). No contexto da genômica de animais de produção, o colapso de modelos é particularmente complexo, pois pode gerar dados genotípicos sintéticos, por exemplo, que não refletem a diversidade natural de alelos, frequências alélicas ou padrões de desequilíbrio de ligação observado nas populações reais (Saxena & Cao, 2020). Essas limitações comprometem o realismo biológico dos dados, reduzindo sua utilidade em subsequentes análises, como otimização de estratégias de seleção ou para inferência de estrutura populacional.

O colapso de modelos ocorre tipicamente quando a rede geradora identifica um subconjunto restrito de dados que enganam consistentemente a rede discriminadora. Com o tempo, esse comportamento se reforça, já que a rede discriminadora se adapta a esse conjunto limitado, restringindo ainda mais a exploração do espaço latente pela rede geradora (Khanuja & Khanuja, 2021; Metz et al., 2016; Wiatrak et al., 2019). Diversas estratégias foram propostas para mitigar esse problema. A estratégia de utilizar "*minibatches*" permite que o discriminador avalie grupos de amostras geradas coletivamente, penalizando a falta de diversidade e incentivando a rede geradora a produzir dados com maior variação (Salimans et al., 2016). Mais recentemente, arquiteturas com múltiplas redes geradoras treinadas contra uma única rede discriminadora foram desenvolvidas, ampliando a cobertura da distribuição de dados e melhorando a representação de populações que possuem dados heterogêneas (Moghaddam et al., 2023). Essas inovações são especialmente relevantes em genômica de animais de produção, onde modelar com precisão a diversidade genética é essencial.

2.3.3 Avaliação dos dados gerados

Mesmo quando o treinamento atinge estabilidade e evita o colapso de modelos, a avaliação dos dados gerados pelos GAN ainda representa um grande desafio devido à ausência de métricas universais e interpretáveis para avalizar a acurácia

dos dados sintéticos gerados (Yilmaz & Korn, 2024). Diferente de modelos baseados em verossimilhança, os GAN não fornecem medidas explícitas de probabilidade dos dados, o que limita a interpretabilidade dos resultados (Saxena & Cao, 2020). Para a geração de imagens sintéticas, por exemplo, métricas como o Inception Score (Salimans et al., 2016) e a Fréchet Inception Distance (Heusel et al., 2018) são comumente utilizadas para avaliar a semelhança entre imagens reais e sintéticas. No entanto, essas métricas não se aplicam diretamente a dados estruturados, como matrizes genotípicas ou de sequência de DNA, o que exige alternativas específicas para o este tipo de dado. Na genômica, existem algumas alternativas para avaliar a acurácia dos dados gerados divididos aqui em métodos qualitativos e quantitativos:

- Abordagens qualitativas: Baseiam-se em comparações visuais entre dados reais e sintéticos para verificar se padrões biológicos são preservados. A análise de componentes principais pode revelar, por exemplo, se os genótipos sintéticos reproduzem a estrutura populacional observada em conjuntos de dados reais (Diaz-Papkovich et al., 2019; van Waaij et al., 2023). Os gráficos de decaimento do desiquilíbrio de ligação, que acompanham a queda da correlação entre marcadores em função da distância física, permitem comparar estruturas de dependência genômica entre dados reais e sintéticos (Ren et al., 2010; Abo-Ismail et al., 2014). Já os histogramas de frequência alélica fornecem outro indicador essencial, como a semelhança nas distribuições entre dados reais e sintéticos demonstram que os genótipos sintéticos capturam níveis naturais de diversidade genética (Rezaei & Hedayat, 2013).
- Abordagens quantitativas: Empregam métricas estatísticas objetivas. Testes de hipótese, como testes de permutação, *t-test* e o teste de Kolmogorov-Smirnov, avaliam se conjuntos reais e sintéticos podem ser distinguidos estatisticamente um dos outros (Oprisanu et al., 2022). Análises de correlação do desequilíbrio de ligação ou frequências alélicas fornecem evidências adicionais de alinhamento entre dados reais e sintéticos, sendo que altos coeficientes de correlação indicam reprodução eficaz dos padrões genômicos observados em dados reais (Yelmen et al., 2021; Shi et al., 2022).

A combinação dessas abordagens oferece um arcabouço holístico para avaliar a qualidade dos dados genômicos gerados por GAN. Esse equilíbrio entre métricas qualitativas e quantitativas é especialmente importante na genômica de animais de produção, onde os dados sintéticos precisam preservar padrões biologicamente relevantes e, ao mesmo tempo, manter robustez estatística.

2.3.4 Considerações finas sobre os desafios enfrentados

Esses desafios ilustram tanto o potencial quanto a complexidade da aplicação de GAN em pesquisa que envolvem biologia ou genética. Problemas de instabilidade, colapso de modelos e limitações de avaliação da acurácia dos dados gerados ainda são consideradas barreiras importantes para adoção de GAN em larga escala. Na genômica e na pesquisa com animais de produção, tais dificuldades são ainda mais acentuadas, pois os dados sintéticos precisam capturar não apenas plausibilidade estatística, mas também características biológicas essenciais, como distribuições de frequências alélicas, estruturas do desequilíbrio de ligação e diversidade populacionais. Superar essas barreiras exigirá inovação metodológica contínua e a adaptação cuidadosa das estratégias já existentes às propriedades únicas dos dados genômicos. Abordar essas restrições é fundamental para explorar todo o potencial dos GAN na geração de dados sintéticos realistas, interpretáveis e úteis para a pesquisa em animais de produção e em outras áreas de pesquisa.

3 EXEMPLO NO USO DE MODELOS GENERATIVOS

Aproveitando os avanços recentes em modelos generativos, apresentamos um exemplo prático do uso de redes generativas adversariais para sintetizar marcadores SNP (single nucleotide polymorphisms) em bovinos. Em particular, exploramos um modelo generativo baseado em Wasserstein com penalização de gradiente e combinado com componentes principais para reduzir dimensionalidade (PC-WGAN). Nessa abordagem o modelo generativo e trainado em um espaço latente biologicamente informativo e, em seguida, reconstrói marcadores SNP no

espaço original. O objetivo é gerar dados sintéticos que preservem padrões essenciais da estrutura populacional, distribuição de frequências alélicas e desequilíbrio de ligação, viabilizando aplicações em simulação de cenários de melhoramento, avaliação de metodologias e aumento de amostras quando dados reais são limitados.

3.1 Descrição dos dados

Para garantir controle sobre estrutura populacional e a varrição genética, utilizamos dados simulados com o programa QMSim (Sargolzaei & Schenkel, 2009), usando parâmetros similares aos observados em um sistema real de produção e ao do genoma bovino. O processo de simulação consiste, primeiro, na simulação de uma população histórica e, depois, de uma população recente. Para a população histórica, foram simuladas 1.000 gerações com aumento de tamanho de 1.000 para 50.000 indivíduos, seguidas por 2.000 gerações não sobrepostas com decréscimo de 50.000 para 20.000 indivíduos, a fim de criar desequilíbrio de ligação, mutações e deriva genética. Nessa fase histórica, o número de indivíduos de cada sexo foi igual, e utilizou-se união aleatória de gametas. Não se assumiu seleção ou migração na população histórica.

A população recente foi simulada selecionando 40 machos e 1.200 fêmeas (fundadores) da última geração da população histórica. Em seguida, simularam-se quatro gerações com sistema de acasalamento aleatório, um filho por matriz por ano, 50% de prenhezes masculinas e taxas de reposição de 30% para fêmeas e 80% para machos. Para reposição, animais com os maiores valores genéticos estimados foram selecionados como pais das próximas gerações, e aqueles com os menores valores genéticos estimados foram descartados. Os valores genéticos foram estimados utilizando um modelo animal por meio da metodologia "*best linear unbiased prediction* (BLUP)" via equações lineares mistas de Henderson (Henderson, 1975), considerando a variância genética aditiva verdadeira. Foi simulado uma única característica quantitativa com herdabilidade de 0,30, usando variância genética aditiva de 0,30 e variância fenotípica constante de 1,0.

No total, foram simulados 500.000 marcadores SNP bialélicos e 2.000 nucleotídeos de característica quantitativa (QTN) bialélicos ao longo das quatro

gerações da população recente. Os SNP e QTN foram simulados para os 29 cromossomos autossômicos de *Bos taurus*, com comprimento (2.333 cM) idêntico ao genoma bovino real. Os SNP foram distribuídos uniformemente em cada cromossomo, enquanto os QTN foram posicionados aleatoriamente. Tanto SNP quanto QTN apresentaram a frequência do alelo de menor proporção maior que 0,05. Os efeitos dos QTN foram amostrados de uma distribuição gama com parâmetro de forma 0,4 e parâmetro de escala calculado internamente para produzir variância genética de 0,3. Assumiu-se taxa de mutação recorrente de 2,5 x 10^{-5} por loco por geração para SNP e QTN. A simulação foi replicada cinco vezes, e apenas os dados de uma das replicações, selecionada aleatoriamente, foram usados. Adicionalmente, neste exemplo apenas os SNP do cromossomo 1 (37.540 SNP) será utilizado. Os SNP foram codificados como números inteiros: 0 para o genótipo homozigoto de referência (AA), 1 para heterozigoto (AB) e 2 para homozigoto alternativo (BB).

3.2 Fluxo de trabalho e arquitetura do modelo gerativo

Para gerar SNP sintéticos usando o PC-WGAN com penalidade de gradiente, três etapas principais foram seguidas: transformação dos dados, treinamento do modelo e geração de dados sintéticos (Figura 2). Na transformação dos dados, aplicou-se redução de dimensionalidade à matriz SNP ($n \times m$, em que n é o número de amostras e m é o número de SNP) por meio de análise de componentes principais (ACP), para capturar a variabilidade genética presente no conjunto de genótipos e a estrutura populacional (McVean, 2009; Reich et al., 2008). A ACP foi realizada sobre a matriz padronizada de genótipos usando o método da covariância implementado na biblioteca *scikit-learn v0.24* do Python 3.9. O número de componentes principais (CP) retidos (k = 796) foi determinado selecionando-se aqueles que, em conjunto, explicavam 90% da variância total dos dados. Esse limiar garantiu a preservação dos componentes mais informativos, reduzindo significativamente a dimensionalidade dos dados. A matriz de CP (n x k) serviu como entrada para o treinamento do PC-WGAN. Os autovetores e os vetores de médias da ACP foram armazenados para permitir a transformação inversa dos CP gerados de volta ao espaço original de SNP. Para melhorar a convergência durante

o treinamento, os CP foram normalizados de forma que os valores estejam no intervalo [−1, 1] usando a fórmula:

$$CP_{esc} = 2.\frac{CP - min}{max - min} - 1,$$

em que CP_{esc} é o CP escalado para [−1, 1]; CP é o CP original; min é o menor valor daquela coluna (entre todos os indivíduos) e max é o maior valor (na mesma coluna).

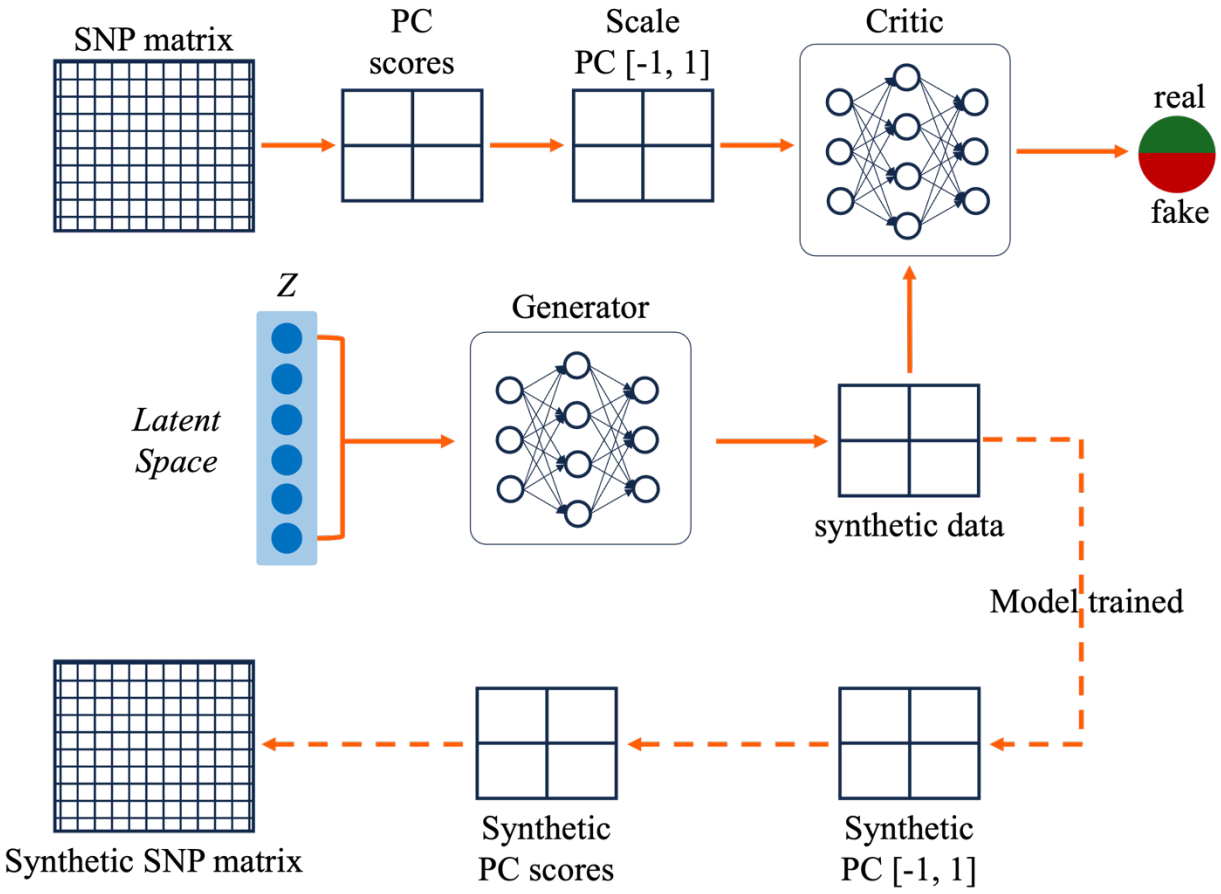


Figure 2. Fluxo de trabalho para a geração de SNP sintéticos usando PC-WGAN. O fluxo se inicia (setas contínuas) com os componentes principais (PC) derivados da matriz de SNP, que são então escalados para o intervalo [−1, 1]. Durante o treinamento, a rede geradora aprende a partir de dados aleatórios de um espaço latente (Z) para produzir PC sintéticos. Esses PC gerados são avaliados pela rede crítica juntamente com os PC reais escalados, permitindo que a rede crítica aprenda a distinguir entre dados reais e sintéticos. Uma vez treinado, a rede geradora pode produzir PC sintéticos a partir de novos dados aleatórios. Conforme indicado pelas setas tracejadas, os PC sintéticos são reescalados para seus valores

originais e transformados inversamente para reconstruir uma matriz de SNP sintética.

A PC-WGAN consiste em duas redes neurais artificiais (Figura 2): a geradora e a crítica (que substitui a discriminadora tradicional na WGAN). O modelo é treinado para minimizar a distância de Wasserstein-1 entre as distribuições de dados reais e sintéticos, com um termo de penalidade de gradiente (λ) adicionado para impor a restrição de continuidade de Lipschitz (Gulrajani et al., 2017). A rede geradora é do tipo "*feedforward*" totalmente conectada, desenhada para dados tabulares. O modelo recebe um vetor Z amostrado de uma distribuição normal padrão, $Z \sim N(0, 1)$, e mapeia para o espaço k-dimensional de CP. A arquitetura contém três camadas ocultas totalmente conectadas, cada uma seguida de função de ativação Leaky ReLU (α = 0,2) e "*batch normalization*" para melhoram a estabilidade durante o treinamento (Khanuja & Khanuja, 2021). O número de neurônios em cada camada oculta foi de 1.024 (camada 1), 512 (camada 2) e 256 (camada 3). A camada de saída recebe 256 valores da última camada oculta e produz k saídas, correspondentes ao número de CP retidos (796). Em seguida, os CP sintéticos passam por uma função de ativação tangente hiperbólica (tanh), compatível com o intervalo [−1, 1] dos CP normalizados.

A rede crítica recebe vetores de CP sintéticos e reais e produz um valor escalar que representa o escore de Wasserstein (Figura 2). A arquitetura da rede crítica espelha exatamente a da rede geradora, com três camadas ocultas totalmente conectadas (1.024, 512, 256 neurônios respectivamente). Cada camada é seguida por uma função de ativação Leaky ReLU (α = 0,2), o que favorece o fluxo de gradiente durante o treinamento (Radford et al., 2016). A camada de saída não possui nenhuma função de ativação e produz um valor que representa quão real são os dados gerado pelo gerador. Aplicou-se normalização espectral a todas as camadas para estabilizar a rede crítica e reforçar a restrição de Lipschitz (Miyato et al., 2018). O modelo PC-WGAN foi implementada em Python 3.9 usando *PyTorch v1.12* (Paszke et al., 2019).

Para o treinamento, utilizou-se uma estratégia de "*holdout*" para dividir os 4.800 individuais em conjunto de treinamento (90%) e teste (10%). A PC-WGAN foi

treinado no conjunto de treinamento com o otimizador Adam tanto para a rede geradora quanto para a crítica, com os parâmetros β_1 = 0,0, β_2 = 0,9 e taxa de aprendizado de 0,0001 para a rede crítica e 0,000005 para a rede geradora. Para satisfazer a restrição de Lipschitz usando a função de perdas baseada em Wasserstein, a rede crítica foi atualizada dez vezes para cada atualização da rede geradora. O termo de penalidade de gradiente na função de perda do PC-WGAN foi definido como 20. Essa penalidade encoraja a norma do gradiente da rede crítica em amostras interpoladas a permanecer próxima de 1, reforçando a restrição de Lipschitz. A PC-WGAN foi treinada por três durações ("*epochs*") diferentes (100, 150 e 500) para avaliar estabilidade e desempenho ao longo do tempo, com "*batch size"* de 64 amostras.

Após o treinamento, a rede geradora foi usada para criar 4.320 amostras sintéticas para comparação com o conjunto de treino e 480 amostras adicionais para compare com o conjunto de teste. Os CP sintéticos gerados foram reescalados para os valores originais dos CP usando a equação:

$$CP_{original} = \frac{(CP_{esc} + 1)}{2}.(max - \min) + min,$$

em que $CP_{original}$ é o CP no domínio original, CP_{esc} é o CP gerado pelo modelo no intervalo $[-1, 1]$, min é o menor valor por coluna de CP e max é o maior valor por coluna de CP. Em seguida, os CP sintéticos foram transformados de volta em SNP usando os autovetores obtidos da ACP nos sobre os dados reais, pela equação (Pearson, 1901; Pedregosa et al., 2018):

$$\hat{X} = Z.W + \mu,$$

em que $\hat{X}$ é o SNP (sintético) reconstruído, Z são os valores dos CP gerados pelo gerador do PC-WGAN, W é a matriz de ACP com os autovetores (matriz de CP) e μ é o vetor de médias calculado a partir dos SNP na etapa de pré-processamento. Por fim, os valores contínuos foram transformados para a classe genotípica mais próxima por arredondamento baseado em limiar para se ter apenas valores 0, 1 ou 2 que representam os SNP.

3.3 Métricas de avaliação

Para avaliar se os SNP sintéticos gerados pelo modelo PC-WGAN capturam adequadamente as propriedades estatísticas e biológicas dos SNP reais de bovinos, empregamos uma combinação de métricas quantitativas e qualitativas que focam em três atributos genômicas chaves, sendo eles: distribuições de frequências alélicas, padrões de desequilíbrio de ligação e estrutura populacional. Em primeiro lugar, avaliamos a distribuição da frequência dos alelos nos dados reais e sintéticos, calculadas a partir da frequência do alelo de menor proporção na população. A similaridade entre as distribuições foi avaliada por correlação de Pearson (r) entre os vetores de frequência dos alelos de menor proporção nos dados real e sintético; correlações altas indicam concordância das distribuições de frequências. Além disso, as distribuições foram comparadas visualmente (histogramas) e testadas quantitativamente usando o teste de Kolmogorov-Smirnov: p-valores não significativos ($p > 0,05$) sugerem que as distribuições são indistinguíveis estatisticamente, apoiando a capacidade do modelo de reproduzir a variabilidade de frequências alélicas.

Na sequência, o padrão no desiquilíbrio de ligação que captura associações não aleatórias entre alelos em locos distintos e essencial para reter a estrutura de haplótipos foi avaliado. O desequilíbrio de ligação foi mensurado como o quadrado da correlação de Pearson (r^2) entre todos os pares de SNP localizados até 100 kb entre si. Para quantificar a preservação do desequilíbrio de ligação, calculamos a correlação entre os vetores de desequilíbrio de ligação dos dados reais e sintéticos. Além disso, curvas de declínio no desequilíbrio de ligação foram geradas utilizando-se o r^2 médio em função da distância física entre pares de SNP. Com isso e possível visualizar como o desiquilíbrio de ligação decai com o aumento da distância entre SNP. Perfis de decaimento sobrepostos e correlações altas entre dados reais e sintéticos ao longo do genoma indicam manutenção de blocos de haplótipos. Um *t-test* pareado foi usado para investigar diferenças significativas nas médias de r^2 em regiões genômicas pareadas; resultados não significativos ($p > 0,05$) indicam reprodução acurada da estrutura de LD.

Por fim, para avaliar a representação da estrutura populacional, análises de componentes principais foram realizadas independentemente nas matrizes de SNP

real e sintética. Os dois primeiros componentes, que explicam a maior parte da variação genética da população, foram extraídos e plotados para visualizar a sobreposição entre as distribuições dos componentes calculados para os dados reais e sintéticos. A visualização dos componentes fornece uma avaliação qualitativa de quão bem a estrutura populacional subjacente foi capturada durante o treinamento do modelo generativo. Para quantificar a similaridade, calculou-se a distância euclidiana entre os centroides de amostras reais e sintéticas no espaço de componentes principais, e aplicou-se um teste de permutação (10.000 iterações) para verificar se a distância observada difere do esperado ao acaso. Quando os p-valores são significativos (p > 0,05) sugerem que a estrutura populacional nos dados sintéticos é consistente com a do conjunto original. Todas as análises computacionais foram conduzidas com *random*, *numpy*, *pandas*, *matplotlib*, *sklearn* e *scipy* no Python 3.9.

3.4 Resultados

O objetivo desse exemplo foi gerar SNP sintéticos que preservassem a estrutura genética subjacente a uma população de bovinos. Foram usados 37.540 SNP do cromossomo 1 e 4.800 individuais; 4.320 amostras foram usadas para o treinamento e 480 para teste. O modelo generativo (PC-WGAN) usado aqui foi inspirado na metodologia proposta por Szatkownik et al. (2024). Em Szatkownik et al. (2024), 4.507 (90% do total) componentes principais (5.008 PCs no total) foram retidos para treinar o modelo a partir de aproximadamente 65.000 SNP. No nosso exemple, apenas componentes principais que capturam 90% da variação genética (796 PCs) derivados de SNP foram retidos para treinar o modelo. Embora o número de componentes principais usados aqui seja cerca de seis vezes menor que em Szatkownik et al. (2024), o modelo capturou de forma eficiente a variação genética e a estrutura populacional subjacente, como mostrado abaixo. Essa redução provavelmente contribuiu para o melhor desempenho computacional: nosso modelo demandou aproximadamente 4 horas para treinar por 500 iterações, enquanto Szatkownik et al. (2024) relataram aproximadamente 20 horas para treinar o modelo por 1.300 iterações. Embora o tempo total também reflita

diferenças de arquitetura, tamanho de conjunto e estratégias de treinamento, os resultados foram satisfatórios e comparável com menor custo e menos iterações.

3.4.1 Estabilidade e convergência do modelo.

Três experimentos com diferentes durações (100, 150 e 500 iterações) foram conduzidos para avaliar estabilidade e convergência (isto é, quando o modelo melhor aproxima a distribuição real). A função de perda do gerador diminuiu gradual e consistentemente ao longo das iterações, aproximando-se de zero nos modelos treinados por 100 e 150 iterações, indicando minimização efetiva da distância de Wasserstein (Figura 3). Já quando o modelo foi treinado por 500 iterações, a função de perda teve um comportamento instável, com oscilações da função de perda do gerador entre 0 e 1 após a iteração 200, sugerindo não convergência ou instabilidade. A função de perda do crítico manteve-se negativa nos três experimentos, como esperado em WGAN (Arjovsky et al., 2017), mas com picos quando o modelo foi treinado por mais de 200 iterações, condizentes com sinais iniciais de não convergência. Essas oscilações na função de perda podem surgir desbalanceamento entre a rede geradora e rede crítica ou fluxo de gradiente insuficiente (desaparecimento de gradiente), levando a resultados pouco confiáveis (Wang et al., 2024). Estudos prévios sugerem que estabilidade pode melhorar com inclusão de ruídos de dados, aumento de diversidade ou regularização adicional de gradiente (Khanuja & Khanuja, 2021). Com base nos resultados, treinar o modelo em uma janela de interações de 120 a 200 pode oferecer melhor equilíbrio entre estabilidade e qualidade.

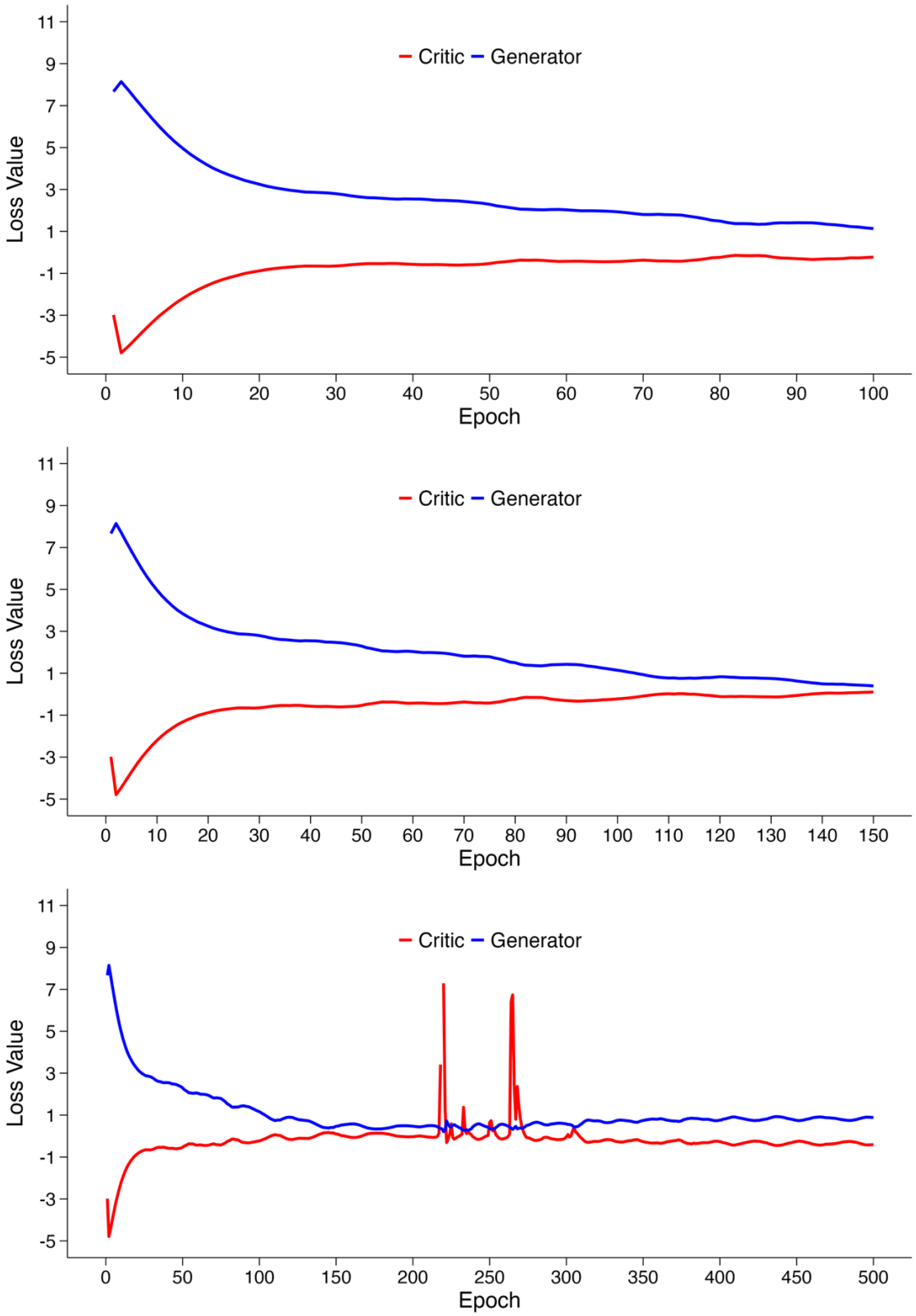


Figure 3. Comparando a dinâmica da função de perda da rede geradora e critica usando três durações para o treinamento do PC-WGAN.

O modelo PC-WGAN treinado aqui não apresentou sinais de desaparecimento de gradiente em nenhum experimento. Essa robustez provavelmente decorre da combinação da arquitetura e de pré-processamento usado: componentes principais escalados para estarem entre −1 e 1, uso de função de ativação Leaky ReLU nas camadas internas da rede geradora, permitindo gradiente pequeno ≥ 0,01 mesmo para entradas negativas (Maas et al., 2013) e função de ativação tanh na saída (mapeando para −1 e 1), o que favorece o modelo a aprender toda a distribuição dos dados (Goodfellow et al., 2014). Resultados similares de estabilidade e convergência foram observados por Szatkownik et al. (2024) ao gerar dados genômicos em humanos com WGAN-GP (Gulrajani et al., 2017), com arquitetura similar a da adotada aqui. Em síntese, o modelo treinado por 150 iterações com 796 componentes principais apresentou o melhor desempenho neste exemplo e, portanto, os resultados a seguir baseiam-se nele. A partir da rede geradora treinada, foram produzidas 4.320 amostras com 37.540 SNP sintéticos cada (espelhando os dados usados no treinamento) e 480 amostras adicionais (espelhando os dados utilizados como teste).

3.4.2 Componentes principais

Para avaliar se a estrutura populacional foi preservada nos dados sintéticos, o primeiro e o segundo componentes principais foram calculados separadamente para as matrizes de SNP reais e sintéticos, e suas distribuições foram comparadas visualmente. Os SNP sintéticos gerados pela PC-WGAN apresentaram forte sobreposição com os dados reais (Figura 4). Tanto no conjunto de dados do treino quanto no teste, a distribuição de amostras sintéticas alinhou-se aos SNP reais, indicando que o modelo capturou a relação genética entre indivíduos que representa a estrutura populacional. Notou-se, ainda, que os SNP sintéticos ultrapassaram levemente os limites do conjunto observado, sobretudo no conjunto de dados de teste, sugerindo geração de SNP novos (mas biologicamente plausíveis) ausentes nos dados reais. Um teste de permutação usando todos os componentes principais foi aplicado para quantificar a distância entre os componentes principais sintéticos e reais; as distâncias estimadas foram 22,72 para o conjunto de dados de treinamento e 22,08 para o conjunto de dados de teste com p-valor < 0,001,

indicando que não há diferenças estatística entre os dados reais e sintéticos. Visualmente, os dados sintéticos recuperam a estrutura central e os padrões de variação, introduzindo diversidade plausível para futuras análises genômicas.

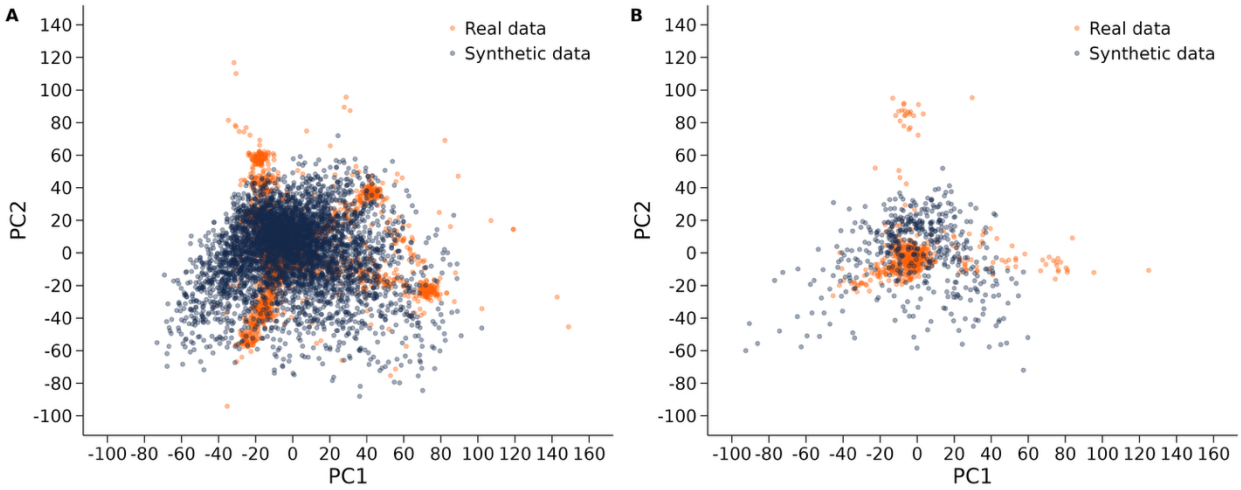


Figure 4. Análise de componentes principais (ACP) para os SNP reais e sintéticos. Gráfico A mostra os ACP para amostras usadas no treinamento e gráfico B mostra o ACP das amostras utilizadas no test. Cada ponto representa uma amostra projetada sobre o primeiro e o segundo componente principal (PC1 e PC2).

3.4.3 Desequilíbrio de ligação

Para avaliar se os SNP sintéticos preservam a estrutura de correlação genômica, comparamos as curvas de decaimento de desequilíbrio de ligação entre conjuntos de dados real e sintético (Figura 5). O modelo reproduziu a tendência geral do desequilíbrio de ligação observado; em curtas distâncias, os valores de desequilíbrio de ligação para os SNP sintéticos ficaram ligeiramente abaixo da observada para os dados reais (subestimando correlação local), enquanto o decaimento se alinhou progressivamente com os valores de desiquilíbrio de ligação dos SNP reais à medida que a distância entre SNP aumentou. As correlações de Pearson entre os vetores de desiquilíbrio de ligação para os dados real e sintético foram 0,94 (conjunto de treinamento) e 0,82 (conjunto de teste). Um *t-test* pareado das médias de r^2 por regiões genômicas resultou em p-valor < 0,001 com diferenças médias pequenas (0,06 em treino e 0,06 em teste), sugerindo semelhança dos padrões de LD. A subestimação em curtas distâncias é consistente com o observado

por Yelmen et al. (2023), que reportaram comportamento semelhante em modelos generativos aplicadas a marcadores SNP em humanos.

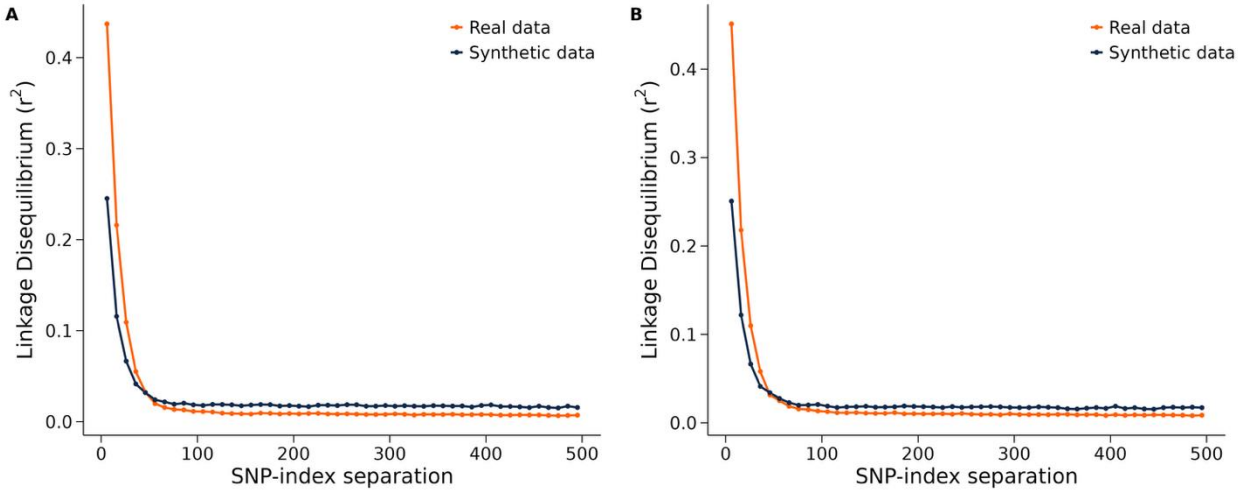


Figure 5. Decaimento no desequilíbrio de ligação para SNP reais e sintéticos para dados do treinamento (A) e teste (B).

3.4.4 Frequência alélica

Para verificar a reprodução do espectro de frequências alélicas, analisamos a distribuição de frequência de alelos de menor proporção na população por meio de métodos qualitativos e quantitativos (Figuras 6). O modelo capturou o formato geral das distribuições reais tanto no conjunto de treinamento quanto no conjunto de teste, com leve superestimação dos alelos mais raros e subestimação dos mais comuns nas extremidades do espectro da distribuição. A regressão ponto-a-ponto entre MAF real e sintética ficou próxima da identidade, indicando aproximação das frequências verdadeiras. As correlações das frequências dos alelos de menor proporção foram 0,72 (conjunto de treinamento) e 0,69 (conjunto de teste). Os testes de Kolmogorov-Smirnov foram significativos ($p < 0,001$) em ambos os conjuntos de dados, com estatísticas muito próximas (0,021 e 0,023), o que indica diferenças sutis, porém detectáveis estatisticamente o que é um efeito esperado com número muito grande de SNP (Ioannidis, 2005). Trabalhos anteriores relataram resultados análogos (Oprisanu et al., 2022; Yelmen et al., 2021), observando que a significância estatística, isoladamente, pode não refletir adequadamente o realismo biológico quando o tamanho amostral é e muito grande.

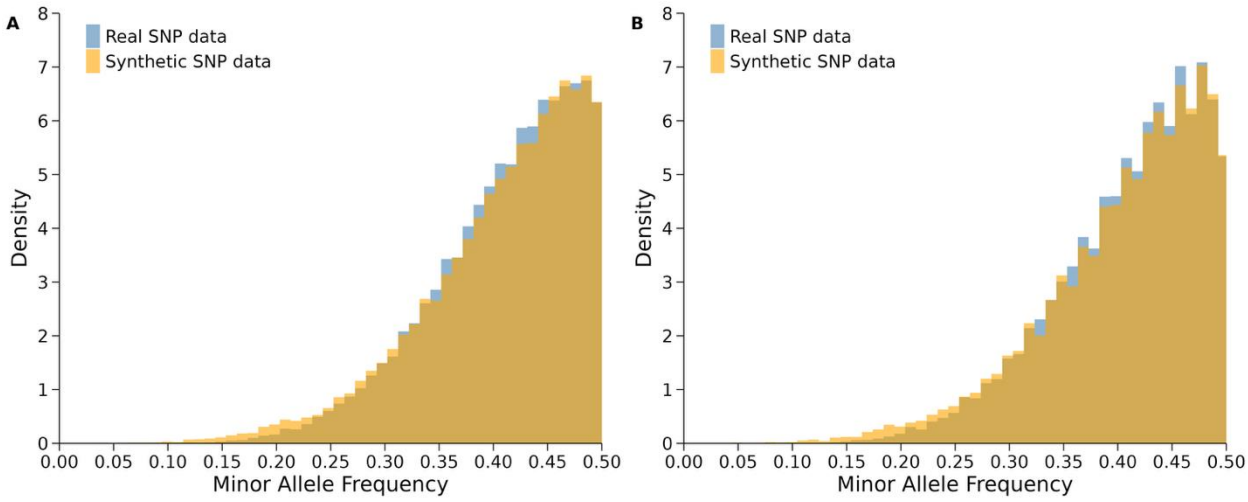


Figure 6. Distribuição das frequências do alelo menor (FAM) para os SNP reais e sintéticos. Gráfico A mostra a distribuição de densidade para a FAM dos dados do treinamento e gráfico B mostra a distribuição dos dados de teste. Os SNP reais estão em azul e sintéticos em laranja.

4 CONSIDERAÇÕES FINAIS

A inteligência artificial generativa representa um campo emergente e em rápida evolução, com potencial de impacto em múltiplas disciplinas pelo seu poder de sintetizar dados complexos e diversos. Entre as arquiteturas disponíveis, as redes generativas adversariais (GAN) têm se destacado pela capacidade de produzir dados altamente realistas – imagens, sequências, textos e dados biológicos – sem replicar exatamente os exemplos do conjunto de dados de treinamento. Há evidências que demonstrando que essas abordagens podem ser aplicadas com sucesso em pesquisas biológicas, especialmente na geração de dados *in silico* para aprimorar modelos estatísticos e de aprendizado de máquina, refinar estudos evolutivos, detectar assinaturas de seleção e solucionar problemas de desbalanceamento de dados.

No contexto da genômica, os GAN já se mostraram úteis para ampliar representatividade em conjuntos de dados, sobretudo em variantes raras ou em fenótipos de baixa herdabilidade, além de possibilitar avanços em tarefas como imputação genotípica, associação genômica ampla, previsão do impacto funcional

de mutações e simulação de respostas celulares a perturbações. Para a pecuária, essas aplicações se tornam particularmente relevantes: dados sintéticos podem complementar fenótipos ou genótipos sub-representados, auxiliar na identificação de biomarcadores e contribuir para estratégias de melhoramento mais eficientes.

O estudo aqui apresentado ilustra essa perspectiva, ao demonstrar a viabilidade de combinar análise de componentes principais com uma GAN baseada em Wasserstein (PC-WGAN) para gerar genótipos sintéticos de SNP em bovinos. O modelo capturou propriedades genômicas centrais, como o decaimento do desequilíbrio de ligação e os padrões de frequências do alelo de menor proporção, além de produzir amostras novas, mas biologicamente plausíveis, mantendo a consistência com a distribuição observada em dados reais. Esses resultados reforçam que modelos generativos adversariais não apenas podem reproduzir estruturas genéticas de populações reais, mas também expandir a diversidade biológica dos dados gerados.

Por fim, os exemplos apresentados na literatura juntamento com o apresentado neste capítulo convergem para uma conclusão clara: os GAN oferecem um caminho promissor como ferramenta complementar em genômica de animais de produção. Ao gerar dados sintéticos realistas, essas redes podem mitigar limitações de acesso a grandes bancos de dados, apoiar análises genômicas subsequentes e, sobretudo, acelerar o desenvolvimento de programas de seleção mais robustos, sustentáveis e orientados por dados. A continuidade desse esforço depende de inovações metodológicas que assegurem a validade biológica das amostras geradas e da integração dessas abordagens em pipelines de pesquisa aplicada, estabelecendo as bases para uma nova era de genômica pecuária orientada por inteligência artificial.

REFERÊNCIAS

Abo-Ismail, M. K., Vander Voort, G., Squires, J. J., Swanson, K. C., Mandell, I. B., Liao, X., Stothard, P., Moore, S., Plastow, G., & Miller, S. P. (2014). Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. BMC Genetics, 15(1), 14. https://doi.org/10.1186/1471-2156-15-14

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. http://arxiv.org/abs/1701.07875

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights, 1(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of Generative Adversarial Networks (GANs): An Updated Review. Archives of Computational Methods in Engineering, 28(2), 525–552. https://doi.org/10.1007/s11831-019-09388-y

Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2023). Video Generative Adversarial Networks: A Review. ACM Computing Surveys, 55(2), 1–25. https://doi.org/10.1145/3487891

Carvalho, T., Belasen, A. M., Toledo, L. F., & James, T. Y. (2024). Coevolution of a generalist pathogen with many hosts: the case of the amphibian chytrid Batrachochytrium dendrobatidis. Current Opinion in Microbiology, 78, 102435. https://doi.org/10.1016/j.mib.2024.102435

Dong, H.-W., & Yang, Y.-H. (2019). On Output Activation Functions for Adversarial Losses: A Theoretical Analysis via Variational Divergence Minimization and An Empirical Study on MNIST Classification. https://arxiv.org/abs/1901.08753

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genetics, 15(11), e1008432. https://doi.org/10.1371/journal.pgen.1008432

Ferreira, A., Li, J., Pomykala, K. L., Kleesiek, J., Alves, V., & Egger, J. (2024). GAN-based generation of realistic 3D volumetric data: A systematic review and taxonomy. Medical Image Analysis, 93, 103100. https://doi.org/10.1016/j.media.2024.103100

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. http://arxiv.org/abs/1406.2661

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved Training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

GM, H., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. Computer Science Review, 38, 100285. https://doi.org/10.1016/j.cosrev.2020.100285

Gangwal, A., & Lavecchia, A. (2024). Unleashing the power of generative AI in drug discovery. Drug Discovery Today, 29(6), 103992. https://doi.org/10.1016/j.drudis.2024.103992

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics, 31(2), 423. https://doi.org/10.2307/2529430

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.

Haque, K. N., Rana, R., Hansen, J. H. L., & Schuller, B. (2020). Guided Generative Adversarial Neural Network for Representation Learning and High Fidelity Audio Generation using Fewer Labelled Audio Data. http://arxiv.org/abs/2003.02836

Hiriyannaiah, S., Srinivas, A. M. D., Shetty, G. K., G.M., S., & Srinivasa, K. G. (2020). A computationally intelligent agent for detecting fake news using generative adversarial networks. In Hybrid Computational Intelligence (pp. 69–96). Elsevier. https://doi.org/10.1016/B978-0-12-818699-2.00004-4

Hazra, D., Kim, M.-R., & Byun, Y.-C. (2022). Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome. International Journal of Molecular Sciences, 23(7), 3701. https://doi.org/10.3390/ijms23073701

Hayawi, K., Shahriar, S., Alashwal, H., & Serhani, M. A. (2024). Generative AI and large language models: A new frontier in reverse vaccinology. Informatics in Medicine Unlocked, 48, 101533. https://doi.org/10.1016/j.imu.2024.101533

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine 19(8): e1004085. https://doi.org/10.1371/journal.pmed.1004085

Iglesias, G., Talavera, E., & Díaz-Álvarez, A. (2024). A survey on GANs for computer vision: Recent research, analysis and taxonomy. https://doi.org/10.1016/j.cosrev.2023.100553

Joo, S., Kim, M. S., Yang, J., & Park, J. (2020). Generative Model for Proposing Drug Candidates Satisfying Anticancer Properties Using a Conditional Variational Autoencoder. ACS Omega, 5(30), 18642–18650. https://doi.org/10.1021/acsomega.0c01149

Khanuja, S. S., & Khanuja, H. K. (2021). GAN Challenges and Optimal Solutions. International Research Journal of Engineering and Technology. www.irjet.net

Li, W., Gauci, M., & Gross, R. (2013). A coevolutionary approach to learn animal behavior through controlled interaction. Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation, 223–230. https://doi.org/10.1145/2463372.2465801

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs Created Equal? A Large-Scale Study. Advances in Neural Information Processing Systems, 31. https://papers.nips.cc/paper_files/paper/2018/hash/e46de7e1bcaaced9a54f1e9d0d2f800d-Abstract.html

Lu, L. (2024). An Empirical Study of WGAN and WGAN-GP for Enhanced Image Generation. Applied and Computational Engineering, 83(1), 103–109. https://doi.org/10.54254/2755-2721/83/2024GLG0066

Liao, S., Lan, S., & Zachariah, A. G. (2024). EVA-GAN: Enhanced Various Audio Generation via Scalable Generative Adversarial Networks. http://arxiv.org/abs/2402.00892

McVean, G. (2009). A genealogical interpretation of principal components analysis. PLoS Genetics, 5(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural NetworkAcoustic Models.

Metz, L., Poole, B., Pfau, D., & Sohl-Dickstein, J. (2016). Unrolled Generative Adversarial Networks.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2017). Least Squares Generative Adversarial Networks. http://arxiv.org/abs/1611.04076

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. http://arxiv.org/abs/1802.05957

Miller, D., Stern, A., & Burstein, D. (2022). Deciphering microbial gene function using natural language processing. Nature Communications, 13(1), 5731. https://doi.org/10.1038/s41467-022-33397-4

Moghaddam, M. M, Boroomand, B., Jalali, M., Zareian, A., Daeijavad, A., Manshaei, M. H., & Krunz, M. (2023). Games of GANs: game-theoretical models for generative adversarial networks. Artificial Intelligence Review, 56(9), 9771–9807. https://doi.org/10.1007/s10462-023-10395-6

Murad, T., Ali, S., & Patterson, M. (2023). Exploring the Potential of GANs in Biological Sequence Analysis. Biology, 12(6), 854. https://doi.org/10.3390/biology12060854

Nayak, R., Pati, U. C., & Das, S. K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106, 104078. https://doi.org/10.1016/j.imavis.2020.104078

Nayak, A. A., Venugopala, P. S., & Ashwini, B. (2024). A Systematic Review on Generative Adversarial Network (GAN): Challenges and Future Directions. Archives of Computational Methods in Engineering, 31(8), 4739–4772. https://doi.org/10.1007/s11831-024-10119-1

Olaniyi, E., Chen, D., Lu, Y., & Huang, Y. (2022). Generative Adversarial Networks for Image Augmentation in Agriculture: A Systematic Review. http://arxiv.org/abs/2204.04707

Okano, R., & Imaizumi, M. (2022). Inference for Projection-Based Wasserstein Distances on Finite Spaces. Statistica Sinica, 34(20), 657–677. https://doi.org/10.5705/ss.202022.0070

Oprisanu, B., Ganev, G., & De Cristofaro, E. (2022). On Utility and Privacy in Synthetic Genomic Data. http://arxiv.org/abs/2102.03314

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018). Scikit-learn: Machine Learning in Python. http://arxiv.org/abs/1201.0490

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. http://arxiv.org/abs/1912.01703

Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M.-C., Rosenstiel, P. C., Teotónio, H., & Schulenburg, H. (2019). The genomic basis of Red Queen dynamics during rapid reciprocal host-pathogen coevolution. Proceedings of the National Academy of Sciences of the United States of America, 116(3), 923–928. https://doi.org/10.1073/pnas.1810402116

Pan, Z., Yu, W., Wang, B., Xie, H., Sheng, V. S., Lei, J., & Kwong, S. (2020). Loss Functions of Generative Adversarial Networks (GANs): Opportunities and Challenges. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(4), 500–522. https://doi.org/10.1109/TETCI.2020.2991774

Qi, G.-J. (2017). Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities. arXiv preprint arXiv:1701.06264.

Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. Nature Genetics, 40(5), 491–492. https://doi.org/10.1038/ng0508-491

Ren, X., Sun, D., Guan, W., Sun, G., & Li, C. (2010). Inheritance and identification of molecular markers associated with a novel dwarfing gene in barley. BMC Genetics, 11(1), 89. https://doi.org/10.1186/1471-2156-11-89

Rezaei, N., & Hedayat, M. (2013). Allele Frequency. In Brenner's Encyclopedia of Genetics (pp. 77–78). Elsevier. https://doi.org/10.1016/B978-0-12-374984-0.00032-2

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. http://arxiv.org/abs/1511.06434

Rannon, E., & Burstein, D. (2025). Leveraging Natural Language Processing to Unravel the Mystery of Life: A Review of NLP Approaches in Genomics, Transcriptomics, and Proteomics.

Schmidhuber, J. (1992). Learning Factorial Codes by Predictability Minimization. Neural Computation, 4(6), 863–879. https://doi.org/10.1162/neco.1992.4.6.863

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5), 680–681. https://doi.org/10.1093/bioinformatics/btp045

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial Training for Free! http://arxiv.org/abs/1904.12843

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. http://arxiv.org/abs/1606.03498

Sinn, M., & Rawat, A. (2017). Non-parametric estimation of Jensen-Shannon Divergence in Generative Adversarial Network training. http://arxiv.org/abs/1705.09199

Saxena, D., & Cao, J. (2020). Generative Adversarial Networks (GANs Survey): Challenges, Solutions, and Future Directions.

Song, J., & Ermon, S. (2020). Bridging the Gap Between f-GANs and Wasserstein GANs. http://arxiv.org/abs/1910.09779

Saqlain, A. S., Fang, F., Ahmad, T., Wang, L., & Abidin, Z. (2021). Evolution and effectiveness of loss functions in generative adversarial networks. China Communications, 18(10), 45–76. https://doi.org/10.23919/JCC.2021.10.004

Szatkownik, A., Furtlehner, C., Charpiat, G., Yelmen, B., & Jay, F. (2023). Towards creating longer genetic sequences with GANs: Generation in principal component space. https://hal.science/hal-04419057

Shi, K., Liu, X., Alrabeiah, M., Guo, X., Lin, J., Liu, H., & Chen, J. (2022). Image Retrieval via Canonical Correlation Analysis and Binary Hypothesis Testing. Information, 13(3), 106. https://doi.org/10.3390/info13030106

Shi, Y., Shang, M., & Qi, Z. (2023). Intelligent layout generation based on deep generative models: A comprehensive survey. Information Fusion, 100, 101940. https://doi.org/10.1016/j.inffus.2023.101940

Szatkownik, A., Furtlehner, C., Charpiat, G., Yelmen, B., & Jay, F. (2024). Latent generative modeling of long genetic sequences with GANs. https://doi.org/10.1101/2024.08.07.607012

Stewart, A., Audet, T., & Pischedda, A. (2025). Intraspecific coevolutionary arms races. In Elsevier eBooks. https://doi.org/10.1016/b978-0-443-15750-9.00098-7

Thirumagal, E., & Saruladha, K. (2021). GAN models in natural language processing and image translation. In Generative Adversarial Networks for Image-to-Image Translation (pp. 17–57). Elsevier. https://doi.org/10.1016/B978-0-12-823519-5.00001-4

Tripathi, S., Augustin, A. I., Dunlop, A., Sukumaran, R., Dheer, S., Zavalny, A., Haslam, O., Austin, T., Donchez, J., Tripathi, P. K., & Kim, E. (2022). Recent advances and application of generative adversarial networks in drug discovery, development, and targeting. Artificial Intelligence in the Life Sciences, 2, 100045. https://doi.org/10.1016/j.ailsci.2022.100045

Tamilmani, G., Devi, V. B., Sujithra, T., Shajin, F. H., & Rajesh, P. (2022). Cancer MiRNA biomarker classification based on Improved Generative Adversarial Network optimized with Mayfly Optimization Algorithm. Biomedical Signal Processing and Control, 75, 103545. https://doi.org/10.1016/j.bspc.2022.103545

Trevisan de Souza, V. L., Marques, B. A. D., Batagelo, H. C., & Gois, J. P. (2023). A review on Generative Adversarial Networks for image generation. Computers & Graphics, 114, 13–25. https://doi.org/10.1016/j.cag.2023.05.010

van Waaij, J., Li, S., Garcia-Erill, G., Albrechtsen, A., & Wiuf, C. (2023). Evaluation of population structure inferred by principal component analysis or the admixture model. GENETICS, 225(2). https://doi.org/10.1093/genetics/iyad157

Wiatrak, M., Albrecht, S. V., & Nystrom, A. (2019). Stabilizing Generative Adversarial Networks: A Survey.

Wolterink, J. M., Kamnitsas, K., Ledig, C., & Išgum, I. (2020). Deep learning: Generative adversarial networks and adversarial methods. In Handbook of Medical Image Computing and Computer Assisted Intervention (pp. 547–574). Elsevier. https://doi.org/10.1016/B978-0-12-816176-0.00028-4

Wolterink, J. M., Mukhopadhyay, A., Leiner, T., Vogl, T. J., Bucher, A. M., & Išgum, I. (2021). Generative Adversarial Networks: A Primer for Radiologists. RadioGraphics, 41(3), 840–857. https://doi.org/10.1148/rg.2021200151

Wenzel, M. (2023). Generative Adversarial Networks and Other Generative Models (pp. 139–192). https://doi.org/10.1007/978-1-0716-3195-9_5

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., & Jay, F. (2021). Creating artificial human genomes using generative neural networks. PLOS Genetics, 17(2), e1009303. https://doi.org/10.1371/journal.pgen.1009303

Yelmen, B., & Jay, F. (2023). An Overview of Deep Generative Models in Functional and Evolutionary Genomics. Annual Review of Biomedical Data Science, 6(1), 173–189. https://doi.org/10.1146/annurev-biodatasci-020722-115651

Yıldız, E., Yüksel, E., & Sevgen, S. (2024). Investigating the effect of loss functions on single-image GAN performance. Journal of Innovative Science and Engineering (JISE), 8(2), 213–225. https://doi.org/10.38088/jise.1497968

Yilmaz, B., & Korn, R. (2024). A Comprehensive guide to Generative Adversarial Networks (GANs) and application to individual electricity demand. Expert Systems with Applications, 250, 123851. https://doi.org/10.1016/j.eswa.2024.123851

SOBRE OS AUTORES

Isabela Gilho Teixeira: Engenheira de Controle e Automação pelo Instituto Federal de São Paulo (IFSP), atualmente cursa pós-graduação em Engenharia e Administração de Sistemas de Banco de Dados pela Unicamp. Possui 4 anos de experiência no SAS, atuando como consultora técnica em projetos estratégicos com foco em dados, análise e implantação de soluções SAS. Tem experiência com clientes de diversos setores, apoiando-os no uso da plataforma SAS em temas como prevenção à lavagem de dinheiro, previsão e planejamento de demanda, modelagem de risco e conformidade (como IFRS9), visualização de dados, implantação e governança de decisões e migração de ambientes.

Sérgio da Costa Côrtes: Estatístico pela Escola Nacional de Ciências Estatísticas (ENCE-IBGE), mestre e doutor em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Atuou por 22 anos como professor na PUC-Rio. Foi coordenador de diversos cursos de graduação, pós-graduação *lato sensu* e de qualificação profissional no Instituto de Educação Superior de Brasília (IESB). Atuou como professor convidado em cursos de EAD na FGV, pós-graduação em Engenharia de Produção na UNESP e pós-graduação em Ciência de Dados na Univille/SC. Foi diretor executivo do IBGE (2004-2011), coordenador interino da ENCE (2006-2011), e Diretor de Tecnologia da Informação da CAPES (2012-2016). Atualmente é consultor em Ciência de Dados, Inteligência Artificial e *Big Data*. Desde 1982 utiliza soluções SAS e em 2024 recebeu o prêmio SAS Educator.

Benjamin Farah: Bacharel em Ciência da Computação e Engenheiro de Telecomunicações pela UnB. Pós-graduado em Gestão Estratégica de TI pela FGV. Possui mais de 12 anos de experiência em projetos de implantação de tecnologias, produtos e soluções SAS nas mais diversas plataformas e nichos de mercado no Brasil e América Latina. Atualmente é *Senior Technical Architect* e *Head* do time TAM (*Technical Account Manager*), provendo serviços especializados para parceiros e clientes do SAS.

Tiago Bresolin: Zootecnista, mestre e doutor em Genética e Melhoramento Animal. Professor no Departamento de Zootecnia da Universidade de Illinois em Urbana-Champaign (EUA). Seu grupo de pesquisa concentra-se na aplicação de métodos estatísticos e biologia computacional em estudos genéticos e genômicos, explorando tecnologias de visão computacional e técnicas de aprendizado de máquina para gerar fenótipos novos e de difícil mensuração, além de implementar aplicações computacionais para coleta, gerenciamento, compartilhamento e segurança de dados.

II Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR)

A análise estatística ocupa um papel central no avanço da ciência e no apoio a decisões estratégicas em diferentes setores da sociedade. Mais do que lidar com números, trata-se de transformar dados em conhecimento, capaz de gerar impacto real em pesquisa, inovação e desenvolvimento tecnológico.

Com esse propósito, a Universidade Federal de Santa Maria promove a 2ª edição do Workshop on Statistical Tools and Analysis for Scientific Research (WSTAR 2025), consolidando-se como um espaço de integração entre a academia, a indústria e a sociedade. O evento busca capacitar acadêmicos, pesquisadores e profissionais no uso de ferramentas estatísticas e de ciência de dados aplicadas à solução de problemas complexos em múltiplas áreas do conhecimento.

A programação desta edição reúne especialistas renomados do Brasil e do exterior, trazendo palestras que vão desde o data storytelling e o uso do SAS® Studio™ na ciência e tecnologia dos alimentos, até temas de fronteira como a arquitetura moderna da Plataforma Viya™ 4 e a geração de dados biológicos sintéticos por modelos generativos. Trata-se de uma oportunidade única de explorar conceitos, técnicas e ferramentas que estão moldando o presente e o futuro da análise de dados.

Ao oferecer um espaço de troca de experiências, aprendizado prático e diálogo interdisciplinar, o WSTAR 2025 reafirma seu compromisso em formar profissionais mais preparados para enfrentar os desafios contemporâneos, atuando de forma crítica, inovadora e colaborativa.

Home Editora
CNPJ: 39.242.488/0002-80
www.homeeditora.com
contato@homeeditora.com
91988165332
Tv. Quintino Bocaiúva, 23011 - Batista Campos, Belém - PA, 66045-315