Construção de dataset dos fatores da produção de leite entre as regiões do Brasil

Max Felipe S. S. Cravo[1]

[1]Instituto de Computação – Universidade Federal do Rio de Janeiro (UFRRJ)
Caixa Postal 68.530 – Rio de Janeiro – RJ – Brazil

Abstract. *Data analysis, often involving data visualization and the application of mathematical models, requires data acquisition and pre-processing. In many contexts, for such data to be properly processed, it must undergo the entire acquisition and pre-processing workflow, ensuring it is in an easily accessible format. Given this context, the present work seeks to examine data concerning factors associated with milk production across the various regions of Brazil, thereby constructing a unified dataset that relates diverse factors directly or indirectly linked to milk production, and finally, to create a visual presentation of this dataset to facilitate its interpretation by a diverse audience.*

Resumo. *A Analise de dados, muitas vezes envolvendo visualização de dados e aplicações de modelos matemáticos, exigem a captação e pré-processamento de dados, em muitos contextos, tais dados para que sejam devidamente processados devem passar por todo o processo de captação e pré-processamento exigindo que estejam em um formato que seja de fácil acesso. Dado tal contexto o presente trabalho busca se debruçar sobre os dados relacionados a fatores ligados a produção de leite nas diversas regiões do Brasil, assim construindo um dataset unificado que relaciona diversos fatores ligados direta ou indiretamente a produção de leite, por fim criar uma apresentação visual deste dataset facilitando sua visualização por diversos leitores.*

1. Introdução

O brasil se vê no cenário global como um grande competidor nos produtos alimentícios advindos do agro, tal setor teve 23,5% de participação no pib brasileiro no ano de 2024, representando um valor expressivo de influência, esse valor agrega tanto os setores pecuários como os setores agrícolas [Confederação da Agricultura Pecuária do Brasil 2025]. Dentro de tal valor, o setor de produção animal, especificamente produção de leite, representa uma porção relevante na geração de valor, indo para além de suprir o mercado interno, segundo a instituição Cepea, estima-se que em 2020 o setor lácteo gerou 77,1 bilhões de reais, o que representa 4% do pib do agronegócio [Grigol 2025]. Tais dados atentam para a importância da manutenção e entendimento da área do setor de laticínios como uma possível área estratégica para o agro brasileiro.

Considerando não apenas a relevância da área de produção de leite como uma área estratégica, mas também os dados disponíveis, surge a possibilidade de se explorar e agregar as informações sobre a produção de leite. Buscando entender tendências e analisar correlações que permitam entender como as diferentes regiões do Brasil se diferenciam

no processo de produção do leite e são ou não afetadas por fatores intrínsecos dado suas características regionais.

Dado tais fatores, realizar a busca e agrupamento de informações sobre a produção de leite e fatores regionais relacionados, poderiam abrir portas para a construção de *data-sets*, especificamente dois, que agrupem informações relevantes para o entendimento de como se situam quanto à produção de leite em diferentes regiões do Brasil.

1.1. Problema

Dentre as regiões mais importantes do Brasil na produção de leite temos a região sul e sudeste, principalmente liderados por estados como: Minas Gerais e Paraná, tais regiões representam 68% da produção de leite no país [Rocha et al. 2020]. Diversas instituições apresentam via WEB dados econômicos sobre a produção do leite nestes estados e nas diferentes regiões do Brasil, porém as diferentes informações como: número de produção, custo e fatores externos como clima, presença de institutos de pesquisa e variações na situação econômica do Brasil não são agregados em tais sites, sendo informações estas distribuídas pela WEB.

Não apenas os dados relacionados a produção do leite, mas também dados relacionados a qualidade do leite, se apresentam dispersos na WEB, mesmo que estejam disponíveis para acesso, estes não se encontram agrupados em formato de processamento.

Tais informações permitem que a produção de leite no brasil seja entendida não apenas pelos seus dados econômicos e ou número de produção total, mas também, permite ter uma visão completa de como diferentes fatores regionais podem influenciar na produção e qualidade final do produto.

Dado esse fator surge a importância da análise de como as diferentes regiões do Brasil se comportam quanto a sua produção de leite, específicamente, permitindo avaliar quais fatores em comuns permitem com que, certas regiões, Sul e Sudeste, se destacam de maneira tão sólida na produção de leite no Brasil [Hott et al. 2019].

1.2. Justificativa

Atualmente 3 principais órgãos, disponibilizam dados sobre fatores econômicos relacionados à produção e venda do leite, dentre eles temos: Centro de Estudos Avançados em Economia Aplicada (Cepea), o Centro de Inteligência do leite (CILeite) este último gerenciado pela embrapa e Instituto Brasileiro de geografia e estatística (IBGE). Todas estas 3 instituições disponibilizam em suas plataformas digitais dados sobre a produção de leite no Brasil, muitas vezes divididos por estados.

Além de tais órgãos que disponibilizam dados economicos da produção de leite o ministério da agricultura juntamente com a Rede Brasileira de Qualidade do Leite (RBQL) trazem informações sobre a qualidade do leite estadualmente e regionalmente.

Com tais informações disponibilizadas, permite com que seja possível a exploração de maneira comparativa sobre os fatores relacionados à produção de leite em diferentes regiões, além de dados estritamente relacionados à produção de leite, seria possível entender quais outros fatores externos e regionais poderiam ainda influenciar esses dados.

Tal disponibilidade crescente nos dados ao mesmo tempo que positivamente permite analisar o estado de cada região, suas características e entender melhor o panorama geral, vem com o problema de armazenar, unificar e gerenciar tais quantidades de dados garantindo sua integridade e permitindo com que sejam devidamente analisadas [Cabrera et al. 2025].

Além da importância de habilitar uma compreensão das diferenças entre regiões a construção de um dataset unificado permite o desenvolvimento de formas de apresentação visual centralizada destes dados, além de permitir a aplicação de métodos de análises estatísticos e baseados em inteligência artificial que permitam entender as correlações entre os dados e regiões além de permitir a construção de métodos preditivos para análise de tendências futuras.

Para que a utilização de um dataset possa ser realizada para inferência sem os riscos associados a veracidades dos dados, é importante que a proveniência dos dados que compõem o dataset estejam disponíveis, principalmente quando se realiza a captura de dados de diversas fontes diferentes, nesses casos a proveniência do pré-processamento e na captura do contexto de captura destes dados garante uma confiabilidade muito maior [Glavic et al. 2013].

Além de fatores que a construção pode trazer para a análise da produção de leite, espera-se que o mesmo possa ser complementado com futuras adições, seja de novos dados ou atualização com novas entradas dos dados existentes. Dados todos os fatores mencionados, o presente trabalho propõe a criação de um dataset unificado de informações regionais referente a produção de leite do brasil levantando fatores de produtividade, econômicos, climáticos e de qualidade, trazendo juntamente a esses dados informações de proveniênciasobre o processo de transformação destes dados, garantindo veracidade no resultado gerado, além disso a construindo formas de visualização intuitiva dos dados facilitando sua leitura para os diversos públicos.

1.3. Objetivo Geral

Coletar dados dispersos pela WEB sobre produção de leite regionais e seus respectivos fatores econômicos, climáticos, de produção de leite do brasil e dados da qualidade do leite nestas regiões, unificando-os, construindo o resultado em formato acessíveis para processamento por outras pesquisas e construindo formas de visualização de tais dados que facilitam seu entendimento para diferentes públicos.

1.4. Objetivos específicos

- Buscar e salvar informações tabulares relacionadas a informações sobre produção, fatores econômicos e climáticos regionais da produção de leite do brasil;
- Buscar e salvar informações tabulares relacionadas a qualidade do leite entre as regiões do brasil;
- Buscar e salvar informações sobre a presença de instituições ligadas à pesquisa e inovação na produção do leite no Brasil;
- Transformar e unificar os dados em um dataset único;
- Criar formas visuais de visualizações do dataset;
- Estabelecer a proveniência dos processos e arquivos gerados;

2. Trabalhos relacionados

Buscando estabelecer uma revisão da literatura existente relacionada ao presente trabalho realizado e possibilitar avaliar as diferenças e enfatizar as contribuições do trabalho, o presente trabalho então buscou trabalhos relacionados à literatura disponíveis em português e inglês.

Utilizando uma string de busca foram buscados em diversos banco de artigos sendo eles: Scielo, SOL, Google scholar, IEEE, MDPI e Research Gate. Em tais banco de artigos foram utilizados as seguintes strings de busca para respectivamente artigos em inglês e português:

- (”data science”OR ”machine learning”OR ”artificial intelligence”OR ”statistical analysis”OR ”predictive modeling”OR ”big data”OR ”data mining”OR ”data analytics”OR ”quantitative analysis”OR ”computational model”) AND (”milk”)
- (”ciência de dados”OR ”aprendizado de máquina”OR ”análise de dados”OR ”modelagem estatística”OR ”analise preditiva”OR ”big data”OR ”mineração de dados”OR ”analise quantitativa”) AND (”leite”OR ”pecuaria”)

As strings foram construídas buscando encontrar artigos que relacionavam processos de análises de dados com a área de produção de leite. Abaixo são apresentados os artigos encontrados em inglês e português.

O artigo “Scenarios for the milk production Chain in Brazil in 2020”, analisa o consumo, produção e tendências na produção do leite porém não aplica uma análise estatística com os dados, focando na abordagem de metodologia de construção de cenários utilizando o método Delphi com especialistas realizando projeções quantitativas [Spers et al. 2013].

O artigo “Estimativa de estro em vacas leiteiras utilizando métodos quantitativos preditivos”, Tem o foco na análise de predição do ciclo de estro de vacas leiteiras da raça Holandesa, buscando utilizar lógica fuzzy e mineração de dados para predição de estro baseado em dados históricos. O foco do artigo é na análise de dados advindos de sensores e em dados sobre a produção de leite das vacas [de Alencar Nääs et al. 2008].

“Milk production systems in Southern Brazil”, o artigo busca a análise da dinâmica, a distribuição espacial, a evolução e a estrutura dos sistemas de produção de leite bovino nas microrregiões no Sul do Brasil, usando dados de 2000 a 2015. Em cima desses dados foram aplicadas Análises de Quociente de Localização (LQ), Análise de Componentes Principais (PCA) e Análise de Cluster [Telles et al. 2020].

“An analysis of Brazilian raw cow milk production systems and environmental product declarations of whole milk”, tal artigo vai na linha de Avaliação do Ciclo de Vida (ACV), com foco em fornecer um ACB de sistemas de produção de leite cru, nos estados do paraná e Minas Gerais e analisar o desempenho ambiental do leite produzido nestes locais com leites produzidos em outras partes do mundo com bases na Declarações Ambientais de Produto (EPDs) válidas [Barros et al. 2022].

“Influence of Meso-Institutions on Milk Supply Chain Performance: A Case Study in Rio Grande Do Sul, Brazil” O estudo analisou a influência das meso-instituições (EMATER) no desempenho das fazendas leiteiras na região da fronteira oeste do Rio

Grande do Sul nos municípios de Alegrete e Sant'Ana do Livramento, especialmente no contexto das Normas Regulamentadoras [Cordeiro et al. 2022].

Quanto às artigos produzidos em países de língua estrangeira, "Big data analytics for empowering milk yield prediction in dairy supply chains" Apresenta a ferramenta Milk Yield Prediction and Analysis Tool (PAT), ferramenta destinada ao produtor de leite, que busca que o mesmo consiga realizar predições sobre sua produção a nível de uma vaca ou grupos de vacas, um dos objetivos do trabalho foi demonstrar como a aplicação de técnicas de análise de dados são úteis na produção de leite.

O trabalho "Comparison of deep learning models for milk production forecasting at national scale", buscou avaliar a aplicação de dez diferentes modelos de deep learning, na predição da produção mensal da leite na França, tais modelos foram testados utilizando dados de entrada climáticos e econômicos, a ideia do trabalho é avaliar se tais modelos obtêm resultados melhores que os modelos de machine learning tradicionais.

O artigo "Machine Learning Approaches for Dairy(Milk) Quality Assurance", foca na utilização de modelos de Machine Learning, avaliando 7 fatores diferentes que podem influenciar na qualidade do leite como: PH, temperatura, gosto, odor, percentual de gordura, cor e turbidez, aplicando tais dados como parâmetros para diferentes algoritmos, a analise demonstrou que odor e turbidez são as características mais determinantes para classificar a qualidade do leite.

"Smart modelling of dairy milk production with machine learning" é um trabalho que foca em testar diferentes algoritmos de inteligência artificial para predizer a produção de leite utilizando dados obtidos localmente pela universidade de Bowen, analisando dados de produção entre 2021 e 2022, testando em 14 modelos de aprendizado supervisionado.

Quanto aos artigos avaliados, Os artigos brasileiros encontrados, que realizam uma análise dos dados de produção de leite no Brasil, possuem focos bem definidos em estados ou municípios específicos, com foco na análise de qualidade do leite e ou para avaliar como pequenos regiões produtores de leite se orientam em suas cadeias de produção de leite.

Percebe-se que a utilização de algoritmos de IA e modelos estatísticos estão associados a definir e identificar características de qualidade na produção do leite e ou em definir diferenças entre regiões e suas respectivas produções.

Não foi identificado um artigo que trabalhe a produção de leite pensando por região brasileira, e ou artigo que trabalhe utilizando dados variados disponíveis sobre as características de cada região quanto a sua capacidade de produção ou fatores relacionados.

Já em relação ao inglês, foi identificado que os artigos estrangeiros encontrados utilizam e realizam testes com diferentes algoritmos de Machine com ênfase em predizer a qualidade do leite e a produção do leite, cabe citar o artigo "big data analytics for empowering milk yield prediction in dairy supply chains"que utilizou ML para auxiliar diretamente o produtor.

Apesar de diferentes artigos buscarem utilizar modelos ML para construção de modelos de predição, não foi encontrado um artigo que realiza comparativamente a

análise entre regiões diferentes, nem analisam características dessas regiões e como influenciam na produção do leite.

Uns dos fatores que podem influenciar na falta de trabalhos estrangeiros que realizam a analise da produção de leite entre regiões seria a já existência de instituições públicas e privadas que já realizam extansa analise desta produção, com uma alta granularidade, foi identificado principalmente duas fontes importantes: CLAL.it[1] que atua nos Estados Unidos e Europa e o *Economic Research Service*[2], atuando apenas nos Estados Unidos, tais serviços estrangeiros já apresentam diversos dados divididos por região com alta granuralidade e correlação de informações.

Percebeu-se dois fatores entre esses trabalhos que foram a falta da disponibilidade de trabalhos que foquem em agrupar informações sobre os dados da produção, trazer a proveniência destes dados e apresentar ao leitor formas de visualizar esses dados de maneira intuitiva.

3. Dataset

A construção dos *datasets* foi realizada em 4 etapas diferentes, a primeira consiste na obtenção dos dados nas suas respectivas origens, a segunda consiste no processamento e limpeza destes dados e a terceira na unificação destes dados e geração do *datasets* e por último o desenvolvimento utilizando a biblioteca prov, da proveniência destes *datasets* e de seus dados.

3.1. Obtenção dos dados

O *dataset* que unifica dados de produção, econômicos e climáticos foi estruturado mediante a compilação de informações de bases de acesso público, focando em fatores econômicos e climáticos associados aos processos de produção e comercialização de leite nas diferentes regiões do Brasil.

As variáveis centrais que compõem o *dataset* incluem: a média de produção de leite (mensurada em milhões); a variação do Índice Nacional de Preços ao Consumidor Amplo (IPCA); a temperatura média anual; a taxa de precipitação média anual; a média do preço do leite pago ao produtor (por litro); e a média do preço do leite comercializado entre produtores.

Os dados foram obtidos de diferentes fontes, seguindo metodologias específicas. Para as variáveis de temperatura e precipitação, utilizou-se o Instituto Nacional de Meteorologia (INMET), através do portal BDMET (https://bdmep.inmet.gov.br/). Os dados foram coletados separadamente para as regiões, compreendendo o período de 2000 a 2023, com granularidade mensal, e englobando todas as estações disponíveis em cada estado. Cabe ressaltar a disparidade observada na cobertura de torres meteorológicas, sendo que a região Sudeste possui um número consideravelmente superior. O processo de aquisição envolveu o registro de e-mail na instituição, o preenchimento de um formulário de requisição de dados (especificando datas e regiões) e, após confirmação, o recebimento dos arquivos em formato CSV por e-mail.

[1]https://www.clal.it/en/index.php?section=produzioni_usa_latte_bovino

[2]https://www.ers.usda.gov/data-products/dairy-data

A variação do IPCA (inflação) foi extraída do Instituto Brasileiro de Geografia e Estatística (IBGE) https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=series-historicas. Dada a dificuldade em localizar esta variação de forma regionalizada, optou-se por utilizar o índice nacional (Brasil), com o objetivo de avaliar como cada região reage à inflação geral. Os dados foram obtidos através da funcionalidade ”exportar”da plataforma, selecionando o formato CSV.

Os valores de média de produção de leite (em milhões) foram provenientes do CILeite https://www.cileite.com.br/leite_numero_producao. Esta fonte já disponibiliza os dados segmentados entre as regiões, permitindo a análise comparativa entre as produções e a influência de outras variáveis. A aquisição ocorreu via link de download direto na página. Similarmente, a média do preço do leite pago ao produtor nas regiões Sul e Sudeste foi obtida no CILeitehttps://www.cileite.com.br/leite_numeros_precos.

Foram utilizados os dados de ”Preços do leite ao produtor no Brasil (deflacionado)”, que ajustam os valores para remover o efeito inflacionário. Para fins metodológicos de generalização em outras regiões (Nordeste e Centro-Oeste), foram utilizados os dados específicos dos estados da Bahia e de Goiás, respectivamente, devido à disponibilidade. A obtenção destes dados também se deu por link de download direto.

Os dados relacionados ao preço do leite comercializado entre os produtores foram obtidos através do CEPEA https://www.cepea.org.br/br/indicador/leite.aspx os dados são disponibilizados por estado, assim são agrupados os dados por estado para facilitar a diferenciação dos dados de cada região.

Com todos os dados salvos de produção, comercialização e climáticos, os mesmos foram importados no script python denominado “script_de_processamento_leite.ipynb”, onde foram transformados em Dataframes, que é um formato similar a uma tabela xsl, que facilita o processamento destes dados, esse processo é realizado utilizando a biblioteca pandas[3]. Durante o processo de obtenção dos dados, foi identificado que para alguns anos, principalmente de 2000 até 2006, não existem dados disponíveis para os seguintes parâmetros:

- média do preço do leite pago ao produtor
- média do preço do leite comercializado entre produtores

para que tais dados ainda pudessem ser utilizados foi utilizado a técnica de imputação de dados, através do método k-nearest neighbors algorithm. levando em conta que os dados não disponíveis eram exatamente os dos primeiros anos 2001 a 2006, optou-se por técnicas que operassem bem em datasets temporais e não dependessem de valores iniciais já definidos como o método de interpolação para eliminação dos dados NaN para garantir um dataset contínuo, que ainda assim garantisse as características de tendências apresentadas, considerando que estamos trabalhando com séries temporais. Essa abordagem foi escolhida tendo em vista os resultados apresentados por [Ahn et al. 2021].

Quanto aos dados relacionados às instituições de pesquisa em produção de leite foi levantado quais as instituições que mais atuam nessa área, foi identificado principalmente duas instituições públicas, a Embrapa[4] e a Rede Brasileira de Qualidade do Leite

[3]https://pypi.org/project/pandas/

[4]https://www.embrapa.br/embrapa-no-brasil

(RBQL)[5], tais instituições atuam diretamente realizando pesquisas e atuando avaliando e orientando sobre a qualidade do leite produzido.

Foram obtidos dados geográficos quanto às localizações dos laboratórios de tais instituições, avaliando como estão dispersos no território brasileiro. Tais dados de localização foram obtidos nos respectivos sites institucionais, onde são apresentadas as cidades onde se situam cada laboratório, dado a cidade são obtidas as informações de longitude e latitude da cidade, todos os dados foram obtidos e transformados de maneira manual.

Já quanto ao *dataset* de qualidade do leite, os mesmos são obtidos através do site https://mapa-indicadores.agricultura.gov.br/publico/extensions/DSN_OQL/DSN_OQL.html, os dados foram obtidos através do acesso a aba "Qualidade do leite", acessando exatamente o serviço para apresentação das séries históricas de qualidade do leite dado pelo título de: "Série histórica UF/Região UF". Essa coleta foi feito no ano de 2025, no mês de dezembro entre os dias 07 e 08.

Os dados são apresentados em gráficos de maneira que para seu acesso foi necessário scrapping manual, realizado pelo autor do artigo, assim para cada região e ano foram obtidos os dados mínimos e máximos dos dados. Os dados históricos disponíveis são de 2013 até 2023, o ano de 2024 só apresenta dados do mês de janeiro e fevereiro.

Um ponto importante, que se caracteriza como uma limitação de disponibilidade de dados é que as regiões do norte e nordeste têm seus dados apresentados como único por região e não por estado, diferentemente das regiões sudeste, sul e centro-oeste. Então não é possível saber exatamente se os dados são a média da região ou de certos estados em específico.

Os dados de qualidade relacionados à análise microbiana são, CSS(Contagem de células Somáticas) - Número de células somáticas presentes no leite. Células somáticas são células de descamação do epitélio da própria glândula mamária e células de defesa (leucócitos) que passam do sangue para o úbere. e CPP(Contagem padrão em placas) - Número de unidades formadoras de colônias de bactérias por mililitro de leite (UFC/ml). Os dados relacionados a análise física de sólidos do leite são, extrato seco total (EST) que é quantidade de todos os componentes do leite, exceto a água. Já o extrato seco desengordurado (ESD) que é a quantidade de todos os componentes do leite, exceto a água e gordura.

Os dados então são obtidos manualmente e transformados no formato CSV agrupados por região/estado e ano. Considerando que a captura dos dados foi feita manualmente e de uma fonte única, o processo de limpeza e de unificação do CSV foi feito manualmante, sem a necessidade de processamento utilizando ferramentas de processamento.

[5]https://www.gov.br/agricultura/pt-br/assuntos/defesa-agropecuaria/laboratorios-credenciados/laboratorios-credenciados/produtos-de-origem-animal/rede-brasileira-de-qualidade-do-leite-rbql

3.2. Limpeza dos dados

O processo de limpeza dos dados, nesse caso dos dados de produção, climáticos e econômicos envolveu a importação dos arquivos csv salvos, em *scripts* python e a realização da leitura dos dados utilizando a biblioteca pandas, o processo de limpeza envolveu, ajustar campos com dados numéricos substituindo a "," por "." para facilitar o processamento destes valores, Além disso como o presente trabalho busca a construção de um dataset temporal, foram definidos as datas de recorte dos dados e para os dados encontrados que apresentavam os dados mensalmente e não uma média do ano, foram agrupados todos os valores de um ano e realizado o cálculo da média desse ano, todos esses processos foram realizados utilizando a biblioteca pandas.

Além destas transformações também foram separados devidamente as informações relacionadas com as diferentes regiões do Brasil. Todos esses processos foram realizados no *script* denominado: "script_de_processamento_leite.ipynb".

3.3. Unificação dos dados

Com os dados limpos, cada dado relacionado a cada parâmetro analisado, foi separado por região e finalmente unidos em um *dataset* único, foi realizado através da função *merge* do pandas onde todos os *dataset* foram sendo unidos em um único, mantendo suas informações originais.

3.4. Proveniência

Após todo o processo de obtenção, limpeza e unificação, foi realizado a construção utilizando a biblioteca prov do python[6], disponibilizada pela W3C. Utilizando tal biblioteca foi elaborado a proveniência dos dados, desde o processo de obtenção, transformação em *dataframe* até a construção do *dataset* final.

Cada parâmetro processado e o próprio *dataset* final possuem suas respectivas proveniências. Tal proveniência gerada é disponibilizada tanto em formato de texto, quanto no formato de gráfico desenvolvido utilizando as próprias funcionalidades da biblioteca prov. Abaixo é apresentado na Figura 1 um exemplo dos grafos de proveniência desenvolvidos para o *dataset* de dados de produção, climáticos e econômicos, já na Figura 2 temos a proveniência do *dataset* de dados de qualidade.

3.5. Informações do ambiente de desenvolvimento

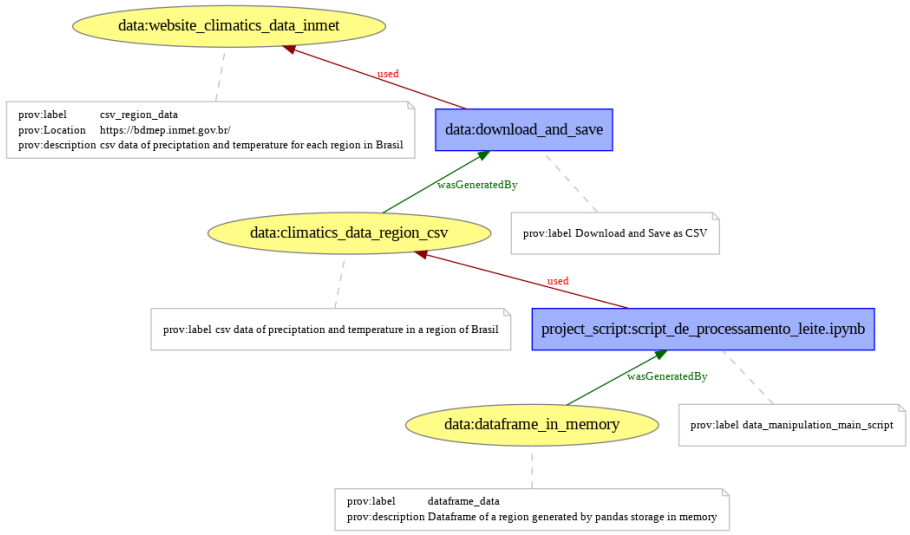
Abaixo são apresentadas as informações sobre o ambiente de desenvolvimento onde foi realizado o processamento dos dados, apresentando as características de hardware e software do ambiente.

3.5.1. Informações sobre o software

Para o desenvolvimento do projeto, a linguagem de programação selecionada foi o Python, em sua versão 3.12.11. O ecossistema técnico foi complementado por um conjunto de bibliotecas essenciais para a análise e visualização de dados, abaixo são apresentadas as bibliotecas utilizadas e suas respectivas versões:

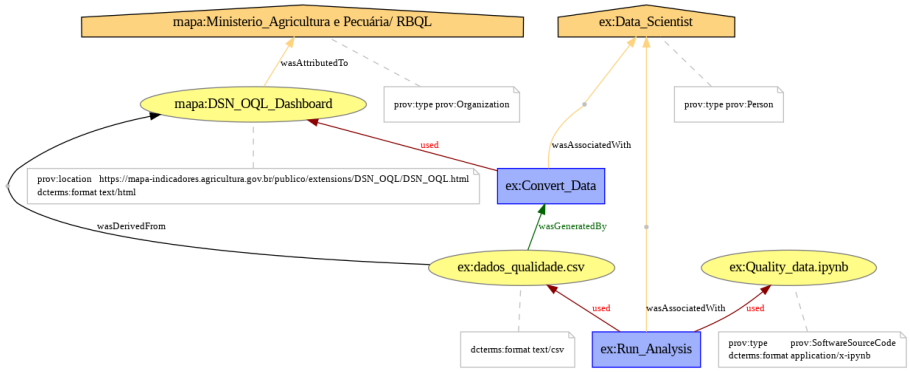
[6]https://pypi.org/project/prov/

Figura 1. Exemplo de grafo de proveniência desenvolvido para o dataset de fatores de produção



Fonte: Elaborado pelo autor

Figura 2. Exemplo de grafo de proveniência desenvolvido para o dataset de qualidade



Fonte: Elaborado pelo autor

- datetime: 3.12.11
- folium: 0.12
- geopandas: 1.1.1
- google: 2.0.3
- graphviz: 0.21
- matplotlib: 3.10.0
- numpy: 2.0.2
- pandas: 2.2.2
- pathlib: 3.12.11
- prov: 2.1.1
- pydot: 4.0.1
- pyvis: 0.3.2
- rdflib: 7.5.0
- requests: 2.32.4

- scipy: 1.16.3
- sklearn: 1.6.1
- statistics: 3.12.11

Uma observação importante, tanto pathlib, datetime e quanto statistics, são bibliotecas built-in do python, ou seja bibliotecas que são instaladas junto com o interpretador, assim relaciona-se a versão de tais bibliotecas com a versão da linguagem na qual está se utilizando.

3.5.2. Informações sobre o Hardware

Quanto ao hardware do ambiente de processamento, cabe citar que os scripts de processamento foram executados no ambiente Google Colab[7], que disponibiliza recursos de hardware para o acesso remoto. Assim obtendo informações do próprio colab quanto ao hardware disponível foram obtidos as seguintes informações:

- Intel(R) Xeon(R)
- 2.20ghz CPU Clock
- 2 núcleos físicos
- 2 núcleos lógicos
- 56320 KB de cache
- Ram: 12.7 GB

Três informações adicionais são importantes de serem mencionadas, O sistema colab não fornece informações quanto a velocidade da memória ram e em nenhum momento da execução dos scripts foi utilizado GPU ou TPU, por fim cabe citar que para auxiliar no armazenamento e recuperação das informações foi utilizado o google drive como armazenamento dos dados, onde os mesmos são acessados diretamente via a biblioteca google disponível nativamente no google Colab.

3.6. Apresentação visual do dataset

Buscando apresentar os dados de maneira concisa e intuitiva foi elaborado uma apresentação utilizando a biblioteca folium, biblioteca essa que permite apresentação e modificação de mapas interativos. Utilizando a biblioteca os dados são apresentados e agrupados por região, onde cada região representada no mapa ao ser selecionada apresenta visualmente uma tabela com os dados daquela região presentes no dataset construído. Acima a Figura 2, apresenta a representação visual do dataset criada utilizando a linguagem python.

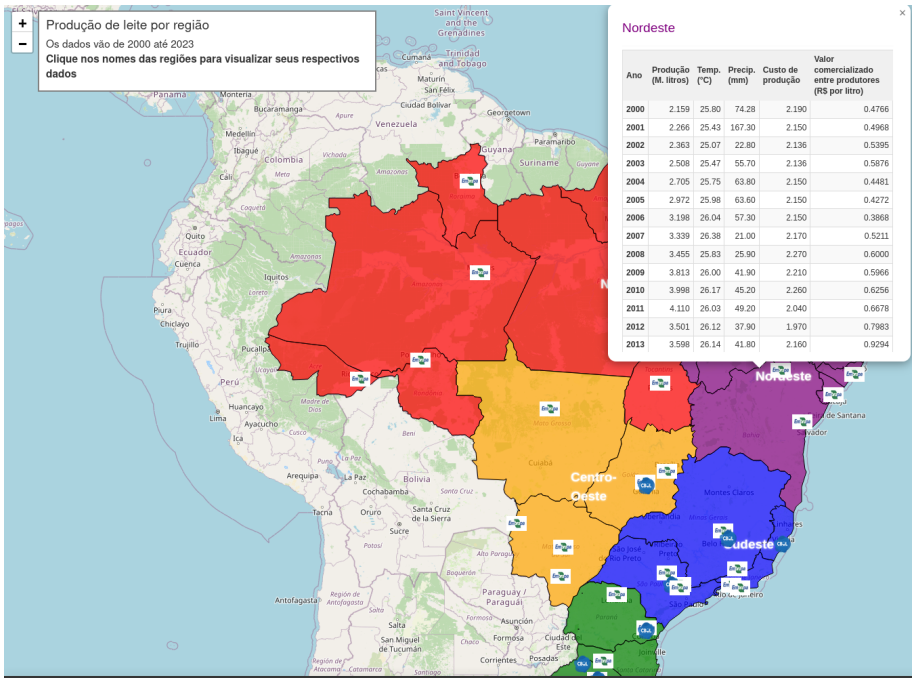
A construção foi constituída utilizando os dados de malhas geográficas disponibilizados pelo IBGE, pela sua API V3, através da biblioteca request do python, especificamente no caso do presente trabalho, os dados de malhas utilizados foram os “intrarregiao”[8].

Além dos dados tabulares agrupados por região são apresentadas no mapa marcadores que apresentam a localização dos laboratórios de pesquisa da Embrapa e da RBQL,

[7]https://colab.google/

[8]https://servicodados.ibge.gov.br/api/v3/malhas/paises/BRformato=application/vnd.geo+json &qualidade=minima&intrarregiao=UF

Figura 3. Apresentação visual do dataset construído



Fonte: Elaborado pelo autor

são marcadas as respectivas cidades sedes de tais laboratórios, além disso, são sinalizados na legenda dos marcadores os laboratórios mais importantes para a produção de leite, que é a de Pelotas que possui uma estrutura que comporta o Sistema de Pesquisa e Desenvolvimento em Pecuária Leiteira (Sispel) e a de Juiz de Fora a Embrapa gado de leite, as duas sendo instituições importantes em pesquisa sobre produção e qualidade do leite.

Essa apresentação visual é inspirada também na apresentação dos dados de qualidade obtidos no site do ministério da agricultura, buscou-se então trazer a mesma forma de apresentação de dados de qualidade para os dados climáticos, produção e econômico.

4. Limitações e conclusões

Em suma, o presente trabalho buscou unificar informações dispersas na rede sobre a produção do leite e seus fatores econômicos, climáticos e qualidade das diversas regiões do Brasil, buscando trazer uma perspectiva geral sobre a produção do leite nas diferentes regiões do Brasil e unificar essas informações em *datasets* que possam ser consultados e ampliados.

A criação dos *datasets* previu também a elaboração da proveniência dos dados, estabelecendo de onde foram obtidos e como foram transformados, bem como uma apresentação visual desses dados, permitindo um entendimento claro das informações das diferentes regiões. A elaboração dos *datasets* e criação da apresentação visual dos dados permitiu com que sejam visualizado alguns pontos importantes sobre a produção do leite e suas diferenças entre as regiões.

Neste contexto apresentar visualmente os dados do *dataset* de produção não é apenas uma etapa do processo de análise de dados que não foi identificado na literatura revisada como também é uma etapa que permite com que os leitores consigam mesmo

sem o interesse de avaliar e manipular o *dataset* em si tenham um incentivo para avaliar o estado da produção de leite per região do Brasil.

Um dos pontos importantes a ser comentado é visualizado que apesar da presença da embrapa nas diferentes regiões do brasil, visualiza-se que a presença de laboratórios e centro de pesquisa em leite se situam predominantemente na região Sul e Sudeste, como exemplo temos, os laboratórios de qualidade do leite (regulamentados) e sedes da embrapa focadas na pesquisa e produção científica do leite como a embrapa gado de leite em Juiz de Fora e o laboratório da embrapa com o Sistema de Pesquisa e Desenvolvimento em Pecuária Leiteira.

(Sispel) que se situa em Pelotas, assim visualiza-se que existem algumas tendências na produção do leite que se relacionam com a presença de unidades públicas de referência focadas na pesquisa e avaliação do leite. Mesmo que a presença de instituições de pesquisa e controle de qualidade não sejam fatores determinantes, visto que tais regiões já são tradicionalmente grandes produtoras de leite, mostram que houve uma centralização geográfica na pesquisa de ponta em leite. Apesar de tal centralização se alinhar com os pontos geográficos mais produtivos, surge a possibilidade se tal centralização não seja detrimental em relação à possibilidade de regiões como a do nordeste crescerem em sua produção.

A região do nordeste apesar de não se situar próxima geograficamente do epicentro da produção de leite teve um aumento significativo de sua produção (191%) mesmo não possuindo tais unidades, tal crescimento pode sugerir que a região tenha potencial a ser explorado podendo vir a ser um novo epicentro da produção de leite, onde neste caso o investimento e descentralização dos setores de análise e desenvolvimento da produção de leite poderiam ser investimento benéficos para tal região.

Por fim, o trabalho apresenta os arquivos e informações construídas disponíveis ao público através da plataforma de compartilhamento de dados Zenodo[9], onde são disponíveis os datasets no formato csv, os scripts e informações de proveniência, a apresentação visual com os dados no formato HTML, que pode ser aberto e visualizado em qualquer navegador e o script utilizado para a geração desta apresentação visual.

4.1. Limitações do trabalho

O trabalho buscou trazer o agrupamento de informações dispersas na rede sobre as diferenças entre as regiões quanto a diversos fatores sobre as suas respectivas indústrias de produção de leite, porém, por falta de tempo e disponibilidade de dados, não adentrando nas especificidades de cada estado e municípios em suas respectivas regiões e definindo como se dão essas diferenças em ambos os datasets, tais informações podem ser importantes para avaliar como estados proeminentes se dão em relação a dispersão da produção estadual.

Outra limitação do presente trabalho se dá na falta de obtenção de informações com relação a variáveis mais detalhadas sobre a produção do leite, como por exemplo número de produção do leite por vaca, número de doenças registradas, outros parâmetros de qualidade do leite, entre outras variáveis que possam auxiliar na identificação de como

[9]https://zenodo.org/records/17714633?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjYwZTYxNzAxLT

tais fatores variam em diferentes regiões.

4.2. Trabalhos futuros

Para trabalhos futuros, seria possível a expansão do dataset criado, adicionando mais parâmetros sobre a qualidade do leite nas diferentes regiões, trazendo mais parâmetros para além de econômicos e climáticos.

Além de expandir os parâmetros, trazer mais granularidade para os dados expandindo o dataset e ou o dividindo para que cada região apresente informações detalhadas sobre a produção entre seus estados e principais municípios traria um melhor entendimento sobre como estados e municípios se diferenciam sobre a produção do leite, além de permitir entender quais municípios se sobressaem quanto a sua produção de leite.

Por fim, trazer inferências utilizando métodos estatísticos e ou baseados em inteligência artificial seria uma forma interessante de utilizar os dados para entender como se dão as tendências de produção de leite nos próximos anos, além disso, permitiria avaliar com mais profundidade se existem relações quanto ao aumento ou diminuição de produção de leite em cada região e como elas se assemelham ou não uma da outra.

Referências

Ahn, H., Sun, K., and Kim, K. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua*, 70(1):767–779.

Barros, M. V. et al. (2022). An analysis of Brazilian raw cow milk production systems and environmental product declarations of whole milk. *Journal of Cleaner Production*, 367:133067.

Cabrera, V. E. et al. (2025). Data integration and analytics in the dairy industry: Challenges and pathways forward. *Animals*, 15(3):329.

Confederação da Agricultura Pecuária do Brasil (2025). PIB do agronegócio registra crescimento de 6,49% no primeiro trimestre de 2025. `https://www.cnabrasil.org.br/`. Acesso em: 13 out. 2025.

Cordeiro, M. P., Viana, J. G. A., and Silveira, V. C. P. (2022). Influence of meso-institutions on milk supply chain performance: A case study in Rio Grande Do Sul, Brazil. *Agriculture*, 12(4):482.

de Alencar Nääs, I. et al. (2008). Estimativa de estro em vacas leiteiras utilizando métodos quantitativos preditivos. *Ciência Rural*, 38:2383–2387.

Glavic, B. et al. (2013). Provenance for data mining. In *TaPP'13*, USA. USENIX Association. Acesso em: 18 nov. 2025.

Grigol, N. (2025). Por que monitorar os preços do leite e dos lácteos? Cepea. Acesso em: 13 out. 2025.

Hott, M. C., Andrade, R. G., and Magalhães Jr., W. C. P. (2019). Distribuição da produção de leite por estados e mesorregiões.

Rocha, D. T., Carvalho, G. R., and Resende, J. C. C. (2020). Cadeia produtiva do leite no Brasil: produção primária. Circular Técnica 123, Embrapa Gado de Leite. Disponível em: `https://tinyurl.com/3y2n2zr9`.

Spers, R. G., Wright, J. T. C., and Amedomar, A. D. A. (2013). Scenarios for the milk production chain in Brazil in 2020. *Revista de Administração*, pages 254–267.

Telles, T. S. et al. (2020). Milk production systems in Southern Brazil. *Anais da Academia Brasileira de Ciências*, 92(1):e20180852.