Construindo um *Dataset* Relacionado à Produção e Comercialização de Produtos da Hortifruticultura no Brasil

Guilherme Alan Mohr[1], Gustavo Pinto da Silva[1], Janaína Balk Brandão[1], Daniel Lichtnow[1]

[1]Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brazil

guialanmohr@gmail.com, gustavo.pinto@ufsm.br, janainabalkbrandao@hotmail.com, daniel.lichtnow@ufsm.br

Abstract. *This paper describes the process of building a dataset that gathers public data related to the production and marketing of horticulture and fruticulture products in Brazil, extracted from various sources using the Web Scraping process. To compose the initial version of the dataset, data was extracted from the 2010 Demographic Census, the Brazilian Institute of Geography and Statistics' (IBGE) Automatic Recovery System (SIDRA), and the National Supply Company (CONAB). Finally, a description of the extracted data and potential use cases is presented*

Resumo. *Este artigo descreve o processo de construção de um dataset que reúne dados públicos relativos à produção e comercialização de produtos da horticultura e fruticultura no Brasil extraídos de diferentes fontes utilizando o processo de Web Scraping. Para compor a versão inicial do dataset, foram extraídos dados do Censo Demográfico de 2010, Sistema IBGE de Recuperação Automática (SIDRA) e da Companhia Nacional de Abastecimento (CONAB). Por fim, é apresentado uma descrição dos dados extraídos e de possíveis usos.*

1. Introdução

Dados relacionados ao agronegócio estão disponíveis publicamente na *Web*, abrangendo áreas como comercialização, produção agrícola e características dos estabelecimentos rurais. Exemplos desses sites são: o Sistema IBGE de Recuperação Automática (SIDRA)[1], Companhia Nacional de Abastecimento (CONAB)[2] e IBGE Censo Demográfico de 2010[3]. Esses dados, são valiosos para pesquisas acadêmicas, permitindo comparações com informações coletadas por outros meios, como pesquisas locais. No entanto, embora esses dados oficiais possam ser consultados, acessados e baixados para análise, o processo muitas vezes envolve realizar múltiplas consultas em interfaces *Web* usando mecanismos de consulta distintos. Iniciativas de reunir dados similares já foram feitas em [Meira, 2002], onde um *Data Warehouse* com dados da produção da fruticultura foi construído a partir de dados obtidos de diversas fontes.

[1] https://sidra.ibge.gov.br/tabela/6954
[2] http://dw.ceasa.gov.br/
[3] https://www.ibge.gov.br/censo2010/apps/sinopse/index.php

Este artigo descreve a construção de um *dataset* abrangendo informações sobre fruticultura e horticultura, obtidas por meio de técnicas de *Web Scraping*. Conforme [Diouf *et al*, 2019], o *Web Scraping*, essencialmente, visa extrair dados de páginas da *Web* para posteriormente integrá-los em um banco de dados, por exemplo.

2. Fontes de Dados Utilizadas e Sistema Desenvolvido

Para a elaboração deste trabalho, em colaboração com os professores do Grupo Interdisciplinar de Pesquisas Agroalimentares Georreferenciadas (GIPAG)[4], foram determinadas as fontes de dados para a versão inicial do *dataset*.

O SIDRA, é um sistema fornecido pelo IBGE, que reúne informações sobre a produção agrícola do Brasil, provenientes de diversas pesquisas, incluindo o Censo Agropecuário de 2017. Os dados estão organizados em 137 páginas distintas, mas com uma interface de consulta uniforme.

A CONAB disponibiliza dados sobre comercialização de aproximadamente 547 produtos nas unidades do Ceasa. O site possui uma interface que permite selecionar medidas, definir colunas e linhas da tabela, além do nível de detalhamento desejado.

Por fim, o IBGE Censo Demográfico de 2010 contém informações sobre a população de cada município e estado, obtidas a partir do censo realizado naquele ano. Para acessar esses dados, deve-se selecionar o estado ou optar pelo Brasil na primeira lista suspensa, e na segunda lista suspensa, escolher a informação desejada.

2.1 Uso das ferramentas de *Web Scraping*

Para implementar o sistema de extração de dados foi utilizado o *Google Collaboratory*, que possibilita criar de *Notebooks* usando a linguagem *Python*. A ferramenta de *Web Scraping* utilizada foi o *Selenium* (versão 4.15), pois para obter os dados, é necessário automatizar ações de arrasto de blocos (*drag and drop*) e *clicks*.

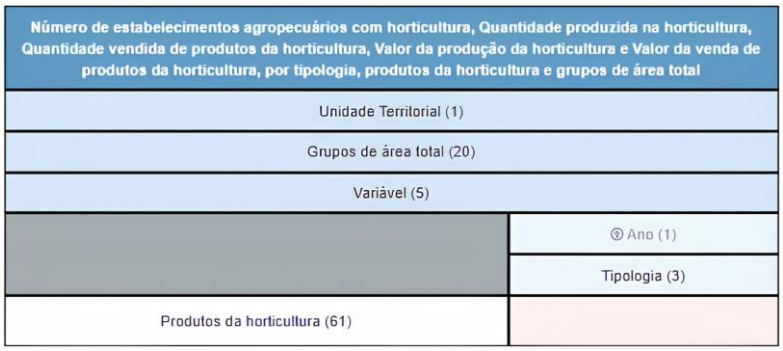

Figura 1. Organizador do layout do retorno dos dados do SIDRA

Na Figura 1, é apresentada a interface que define o layout da tabela final. Para configurar essa organização, utilizou-se o método *drag_and_drop* do *Selenium*. Em seguida, seleciona-se os *checkboxes* adequados para detalhar os dados, permitindo a geração da tabela na mesma página onde foi realizada a consulta inicial.

[4] https://www.ufsm.br/grupos/gipag

Figura 2. Sistema CONAB

No CONAB, o processo de extração envolve a seleção das medidas desejadas, tais como quantidade, valor e preço médio. Em seguida, aplica-se os filtros a fim detalhar os dados (na Figura 2, foram aplicados os filtros de Unidade Federativa (UF) e de ano de comercialização). Posteriormente, seleciona-se a informação que será apresentada nas linhas.

3. Dataset gerado

O *dataset* gerado consiste em dois arquivos em formato CSV, um com dados de produtos da horticultura e outro com dados da fruticultura. Optou-se por utilizar o formato CSV devido à sua ampla aceitação e familiaridade entre o público-alvo, que geralmente não é composto por profissionais da área de Computação, mas sim por pessoas interessadas na análise de mercados de produtos da horticultura e fruticultura.

Na Tabela 1 e na Tabela 2, é apresentado um resumo do *dataset*, visando descrever a abrangência do *dataset* gerado, que está disponível para acesso público em: https://github.com/dlichtnow/GipagDatasetProdCom. Para gerar o *dataset*, foram necessárias 7,5 horas de execução do *script* desenvolvido[5]. Os dados estão organizados a fim de permitir a análise por produto e por estado.

Tabela 1. Quantitativo dos dados presentes no *dataset*

ITEM	Quantidade	
	Fruticultura	Horticultura
Número de registros	1.257	1.404
Número de produtos	46	52

[5] Os processadores são Intel Xeon, modelo 79, operando a 2.20GHz, cada um com 56.320 KB de cache, e pertencem ao mesmo conjunto físico, com 2 núcleos e 1 core cada e a memória possui 13,29 MB.

Tabela 2. Descrição dos principais dados extraídos e presentes no *dataset*

Dado Extraído	Descrição
Nome e Sigla do Estado	IBGE
População Total, Urbana e Rural do Estado	IBGE
Preço médio praticado, Quantidade e Valor total da comercialização do produto na unidade da Ceasa na UF de 2018 até 2022	CONAB
Quantidade produzida e vendida, Valor da produção e da venda do produto no estado	SIDRA
Número de Estabelecimentos dividido em 6 classes: Total, com até 10 ha, de 10 até 100 ha, de 100 até 1000 ha, com mais de 1000 ha e com Produtor sem área	SIDRA

4. Considerações Finais

A base de dados foi construída a partir de discussões entre pesquisadores do GIPAG. Cabe destacar que o GIPAG, construiu bases de dados a partir de levantamentos feitos com produtores da Região Central do RS em visitas as propriedades. Estes dados já foram utilizados em várias publicações [Brandão et al 2023], mas nem todos foram disponibilizados como públicos, uma vez que alguns são dados sensíveis. O foco dos estudos esteve nos mercados acessados por estes produtores. Neste sentido, a construção dessa base visa dar continuidade a análise dos mercados. A partir do *dataset*, uma análise futura consistirá em identificar a demanda de um produto versus sua oferta.

Cabe salientar, que posteriormente pretende-se aprofundar as análises feitas e avaliar os dados extraídos junto a pesquisadores. Uma das dificuldades foi lidar com o nome dos produtos que podem aparecer de forma distinta nas diversas bases. Assim, trabalhos futuros envolvem uso de outras técnicas, como as propostas em [Medeiros & Gonçalves, 2023] que utilizam técnicas de similaridade de *strings* para deduplicação. Futuramente dados de censos mais recentes deverão ser incluídos (esta é inclusive uma limitação do trabalho, visto que o último censo agrícola é de 2017), bem como dados de outras bases de dados, visando gerar bases em outros formatos (e.g. relacional).

Agradecimentos. Trabalho apoiado pelo Edital Conjunto de Circulação Interna do Colégio Politécnico da UFSM.

Referências

Brandão, J. B. et al. (2023) Mercados e canais de comercialização na região central do RS: fatores relevantes para os produtores de frutas e hortaliças. *Ciência Rural*, *53*

Diouf, Rabiyatou et al. (2019) Web scraping: state-of-the-art and areas of application. In: IEEE International Conference on Big Data (Big Data). IEEE. p. 6040-6042.

Medeiros, A. M. A., Gonçalves, E. C. (2023) Estudo Comparativo de Estratégias para o Pareamento de Nomes de Entidades na Língua Portuguesa. In: Anais XVIII ERBD.

Meira, C. A. A. et al. (2002) Análise da produção brasileira de frutas a partir do armazém de dados da fruticultura. Campinas, SP: Embrapa. 6 p. Disponível em: http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/8617. Acesso em: jun/23