**Aluno: Sergio Silvestri Silva**

**Orientadora: Prof.ª Gabrielle Maria Romeiro Lombardi**

**Curso: MBA em Data Science e Analytics**

**Analisar Emissões de Gases do Efeito Estufa por Consumo Familiar Aplicando Modelos de Machine Learning**

**Introdução**

A atividade humana tem contribuído significativamente para o aquecimento global por meio da emissão de gases de efeito estufa (GEE). Esses gases, naturais ou resultantes de atividades humanas, absorvem e emitem radiação em comprimentos de onda específicos, provocando o efeito estufa. Os principais GEE na atmosfera incluem vapor de água (H2O), dióxido de carbono (CO2), óxido nitroso (N2O), metano (CH4) e ozônio (O3). Além disso, gases produzidos exclusivamente por atividades humanas, como hexafluoreto de enxofre (SF6), hidrofluorcarbonos (HFCs), clorofluorcarbonos (CFCs) e perfluorcarbonos (PFCs), também contribuem para esse fenômeno. As emissões de GEE são frequentemente expressas em equivalentes de CO2 (CO2-eq) usando o Potencial de Aquecimento Global (GWP100) para padronizar sua mensuração (IPCC, 2023).

As projeções indicam que, sem medidas adicionais, o aquecimento global poderá ultrapassar 1,5 °C durante o século XXI, dificultando a meta de limitar o aumento abaixo de 2 °C até 2030. Entre 2011 e 2020, a temperatura da superfície terrestre já havia atingido um aumento de 1,1 °C em relação ao período de 1850-1900, conforme o Sexto Relatório de Avaliação (AR6) do Painel Intergovernamental sobre Mudanças Climáticas (IPCC, 2023).

O crescimento das emissões de GEE está relacionado ao uso insustentável de energia, mudanças no uso do solo, padrões de consumo e desigualdade na distribuição dos recursos entre regiões, países e indivíduos. Em 2019, as emissões per capita de GEE associadas ao consumo familiar variaram de 2,6 tCO2-eq a 19 tCO2-eq, com uma média global de 6,9 tCO2-eq. Os 10% da população com maior emissão foram responsáveis por 34% a 45% das emissões globais. Na Europa, a média per capita foi de 7,8 tCO2-eq, representando 8% das emissões globais (IPCC, 2023).

A União Europeia tem priorizado a redução das emissões de GEE, e as Emissões Baseadas no Consumo (EBC) representam um desafio crítico para a diminuição da pegada de carbono da região (Axelsson, 2024). A Comissão Europeia desenvolveu uma estrutura de avaliação baseada na Avaliação do Ciclo de Vida (ACV) para monitorar os impactos ambientais do consumo nas áreas de alimentação, transporte, moradia, bens domésticos e eletrodomésticos (European Commission, 2021). Inclusive, a Plataforma Europeia de Avaliação do Ciclo de Vida (EPLCA) fornece uma base de conhecimento sobre produção e consumo sustentável na UE, na qual pode-se observar que em 2022, os 27 Estados-Membros da UE (UE-27) emitiram, em média, 8,2 tCO2-eq per capita, sendo 38,2% na alimentação, 23,6% no transporte, 21,4% na moradia, 13,7% em bens domésticos e 3,1% em eletrodomésticos. Entre os produtos alimentares, a carne suína (18,4%), queijo (13,7%), aves (9,7%), carne bovina (9,5%) e manteiga (5,2%) foram os maiores contribuintes para as emissões de GEE (European Commission, 2013).

Diante desse contexto, esta pesquisa aplicará modelos de Machine Learning para analisar a pegada de consumo na área de alimentação de uma família de imigrantes brasileiros, composta por quatro membros, residente na República Tcheca. Os dados, referentes ao período de 2018 a 2024, serão analisados segundo a abordagem da Pegada de Consumo da EPLCA. O estudo projetará o consumo até 2030 e avaliará se as emissões de GEE aumentarão, comprometendo as metas de redução, ou se haverá uma diminuição. Caso haja aumento, serão identificadas alternativas de consumo sustentáveis que equilibrem impacto ambiental, nutrição e custo de vida da família.

**Objetivo**

Identificar os melhores modelos de Machine Learning que possam ser aplicados para analisar a pegada de consumo na área de alimentação, para alimentos comprados em supermercados, no período 2018 a 2024, de uma família de imigrantes brasileiros de quatro membros, morando na cidade de Brno na República Tcheca, usando como base de comparação a abordagem de Pegada de Consumo da EPLCA.

Os modelos identificados de Machine Learning serão aplicados para responder às seguintes perguntas:

1. Quais são as emissões de gases do efeito estufa baseadas no comportamento de consumo familiar e como se comparam com a UE-27 e República Tcheca?
2. Qual é a previsão de emissão de gases do efeito estufa per capita em 2030, se continuar com esse comportamento de consumo?
   1. Se houver aumento nas emissões de GEE, quais seriam os alimentos que podem ajudar na redução de emissões até 2030, de modo a não comprometer a saúde nutricional e o custo de vida da família?

**Metodologia**

A pesquisa seguirá a hierarquia entre dados, informação e conhecimento, conforme explica Favero (2024). Os dados tratados e analisados, transformam-se em informação, e o conhecimento é gerado quando as informações são reconhecidas e aplicadas na tomada de decisão.

As perguntas a serem respondidas exigirão uma detalhada análise da base de dados, utilizando técnicas estatísticas e de modelagem multinível, explicado em Favero (2024).

Identificar o modelo de Machine Learning apropriado para a tomada de decisão, exigirá a identificação de Técnicas Multivariadas Exploratórias (análise de agrupamento, análise fatorial por componentes principais e análise de correspondência) e Confirmatórias (modelos de regressão linear e não linear, modelos logísticos de classificação binárias e multinomiais, modelos de contagem, modelos de dados de painel) (Favero, 2024), bem como, modelos de Séries Temporais e Árvores de Decisão para analisar os padrões, tendências, sazonalidade, testar hipóteses e validar as suposições sobre os dados.

As ferramentas de programação e processamento dos dados utilizados na pesquisa serão a linguagem de programação Python, bibliotecas como Pandas, NumPy, Matplotlib, entre outros, ambiente de desenvolvimento integrado Spyder e interface gráfica de usuário Anaconda Navigator.

O recurso a ser usado para a coleta dos dados será o Levantamento de Dados Secundários, para construção das análises, dos resultados e discussão. Conforme explica Malhotra (2004), os dados podem ser coletados e acessados a partir de bancos de dados em que os dados já passaram pelo processo primário de coleta e tratamento de dados, facilitando ao pesquisador o acesso às informações e as análises necessárias. Segundo Malhotra (2004), os dados secundários são classificados como dados Internos onde os dados e informações são oriundos dentro da empresa, instituição, escola, organização em que a pesquisa se encontra o processo de investigação, e dados externos que são aqueles provenientes por meio de publicações em bases de dados, sejam eles artigos, relatórios de pesquisa, relatórios censitários, dados de Governos, instituições sem fins lucrativos, guias, anuários, dados estatísticos, dados de censos, dentre outros.

Os dados internos serão os dados de consumo de compras de alimentos da família de imigrantes brasileiros, com cada observação sendo os itens das notas fiscais de compras. As compras são feitas em supermercados físicos e online. Os supermercados de loja física oferecem um cartão fidelidade virtual através de aplicativo no telefone celular. Cada vez que o cartão fidelidade virtual é lido na caixa registradora, os itens comprados ficam registrados. Os supermercados online registram as compras no site. Em ambos os meios de compras, as notas fiscais podem ser baixadas em arquivos PDF (portable document format em inglês). Esse formato é usado quando é necessário salvar arquivos que não podem ser modificados, mas ainda precisam ser facilmente compartilhados e impressos.

Os dados externos serão os dados para analisar as emissões de GEE adotando a unidade de medida CO2e, sendo usados dados provenientes de banco de dados de um relatório de pesquisa e uma empresa especializada. Para identificar as informações nutricionais dos itens comprados, serão usadas informações nutricionais de uma instituição de pesquisa.

Cada banco de dados coletado será armazenado como uma planilha em um arquivo Microsoft Excel com o nome consumo.xlsx para facilitar a visualização e manipulação.

O banco de dados interno será chamado nota\_fiscal. Para cada layout de nota fiscal em PDF, será criado um script Python para ler o arquivo PDF e gravar um arquivo CSV (comma-separeted values em inglês). Um arquivo CSV é um formato de arquivo de texto que usa vírgulas para separar os valores, e quebras de linha para separar os registros. Armazena dados tabulares em texto simples, onde cada linha do arquivo geralmente representa um registro de dados. A primeira linha do arquivo CSV conterá os nomes das colunas supermercado, loja, data, item, quantidade, preco\_unitario e preco\_total. A partir da segunda linha, conterá as informações de cada item da nota fiscal. Após a conversão para PDF, serão aplicadas as técnicas de Data Wrangling. Alguns layouts usam vírgula como separador de decimais e outros usam ponto. O código padronizará para ponto. Será escrito um código para consolidar os arquivos CSV em um único arquivo com o nome nota\_fiscal.csv. Serão criados códigos para analisar a existência de informações duplicadas e eliminá-las. Serão criados códigos para identificar informações faltantes e completá-las automaticamente ou manualmente. Será criado um arquivo Microsoft Excel com o nome consumo.xlsx. O arquivo nota\_fiscal.csv será carregado no arquivo consumo.xlsx com o nome da planilha nota\_fiscal.

O primeiro banco de dados externo se chamará item e será criado usando um script Python para ler a planilha nota\_fiscal do arquivo consumo.xlsx e armazenar em um dataframe do Python chamado nota\_fiscal. Serão selecionados os itens únicos da coluna item do dataframe nota\_fiscal, classificados em ordem alfabética, e salvos em um arquivo CSV com o nome item.csv e uma coluna chamada item. O arquivo item.csv será carregado dentro do arquivo consumo.xlsx com o nome da planilha item. Serão criadas 6 colunas adicionais na planilha item. A estimativa de CO2e dos alimentos é calculada por quilograma (kg) ou litro (l) do produto. Alguns itens incluem a unidade de medida e a medida do volume na descrição. Para aqueles que não incluem, será pesquisado no site da loja para identificar os valores. A coluna unidade\_medida será preenchida manualmente através de identificação visual da descrição do item ou através de consulta no supermercado onde o produto foi comprado. A coluna medida será padronizada para quilograma (kg) ou litro (l) do produto. Por exemplo, 180g para 0.180kg e 500ml para 0.500l. Para comparar as emissões da família com as emissões por pessoa da UE-27 e República Tcheca, será usada a metodologia defina na plataforma EPLCA. A coluna area\_consumo terá as categorias alimento e não-alimento, pois a pesquisa será analisar as emissões de GEE de alimentos. A coluna grupo terá as mesmas categorias de grupos de produtos da plataforma EPLCA. Os nomes dos produtos serão semelhantes aos produtos de cada categoria de grupo de produtos, conforme definido na plataforma EPLCA. As notas fiscais são emitidas em inglês e tcheco. Cada item será traduzido para o português usando o Google Translate para identificar o nome do produto. Quando a descrição for muito resumida, será pesquisado no site do supermercado para identificar o produto. A República Tcheca importa muitos produtos, tais como, carne bovina, tomate, açúcar, frutas. Essa informação é necessária para estimar as emissões de GEE geradas pelo transporte. A coluna origem terá o nome do país de origem do produto. Será pesquisado no site do supermercado para identificar o país de origem. Para os produtos não alimentícios, a informação não será incluída.

O segundo banco de dados externo se chamará co2e\_produto. Será criado um script Python para ler a planilha item do arquivo consumo.xlsx, selecionar a categoria alimento da coluna area\_consumo, criar um dataframe com os valores únicos da coluna produto, classificar em ordem alfabética e salvar em um arquivo CSV chamado co2e\_produto.csv com uma coluna chamada produto. O arquivo co2e\_produto.csv será carregado dentro do arquivo consumo.xlsx com o nome da planilha co2e\_produto. Será criada uma coluna adicional na planilha co2e\_produto com o nome co2e\_produto. Para estimar o co2e do produto, serão usados os dados do banco de dados aaq0216\_datas2.xls da pesquisa Reduzir os Impactos Ambientais dos Alimentos Através dos Produtores e Consumidores, planilha Results - Retail Weight, Coluna K - GHG Emissions (kg CO2eq/FU, IPCC 2013 incl. CC feedbacks), Median (Poore, 2018). Quando a base de dados não tiver o dado do produto, serão utilizados os dados do site Carbon Cloud (Carbon Cloud, 2019).

O terceiro banco de dados externo se chamará co2e\_transporte. Será criado um script Python para ler a planilha item do arquivo consumo.xlsx, selecionar a categoria alimento da coluna area\_consumo', criar um dataframe com os valores únicos da coluna origem, classificar em ordem alfabética e salvar em um arquivo CSV chamado origem.csv com uma coluna chamada origem. Será criada uma coluna adicional na planilha co2e\_produto com o nome co2e\_transporte. Para estimar o CO2e emitido pelo transporte, serão usados os dados do site CO2 Everything e a distância estimada da origem até o destino (CO2 Everything, 2019). Produtos vindos de outros continentes chegam de navio no porto de Rotterdam na Holanda e seguem de caminhão até o supermercado. Produtos produzidos na Europa são transportados de caminhão. As distâncias serão estimadas baseadas em pesquisas na Internet.

O quarto banco de dados externo se chamará caloria. Dentro do arquivo consumo.xlsx, será criada uma planilha com o nome caloria e copiada a coluna produto da planilha co2e\_produto para a planilha caloria. Serão criadas 5 colunas com os nomes proteina, gordura, carbo, fibra e alcool. Para cada produto, as informações serão pesquisadas no banco de dados da Tabela Brasileira de Composição de Alimentos (TBCA, 2013). Quando o produto não for encontrado na tabela, serão usadas as informações do rótulo do produto ou do site do supermercado. Será criada uma coluna adicional na planilha com o nome caloria\_total\_produto para calcular a quantidade de calorias em kcal (quilocalorias) por 100g ou 100/ml usando a fórmula caloria\_total\_produto = (proteína x 4) + (gordura x 9) + (carbo x 4) + (fibra x 2) + (álcool x 7).

O quinto banco de dados se chamará consumo com a consolidação de todos os bancos de dados. Serão criadas as colunas co2e\_item = quantidade x medida x (co2e\_produto + co2e\_transporte) e caloria\_item = quantidade x medida x caloria\_total\_produto x 10 (a multiplicação por 10 é necessária para obter o valor da caloria em quilograma ou litro).

Será criada uma planilha com o nome dicionario\_de\_dados com o detalhe de cada banco de dados e colunas.

**Resultados Esperados**

Espera-se identificar os melhores modelos de Machine Learning para analisar o comportamento de consumo da família de imigrantes brasileiros e responder as perguntas definidas nos objetivos.

**Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Definir do projeto de pesquisa | X |  |  |  |  |  |  |  |  |  |
| Preencher o Formulário de Direcionamento Ético (FDE) | X |  |  |  |  |  |  |  |  |  |
| Reunir com a orientadora | X | X |  |  |  |  |  |  |  |  |
| Enviar o projeto de pesquisa para a orientadora |  | X |  |  |  |  |  |  |  |  |
| Ajustar o projeto de pesquisa |  | X |  |  |  |  |  |  |  |  |
| Entregar o projeto de pesquisa |  | X |  |  |  |  |  |  |  |  |
| Coletar os dados | X | X |  |  |  |  |  |  |  |  |
| Analisar os modelos de Machine Learning |  | X | X |  |  |  |  |  |  |  |
| Reunir com a orientadora |  | X | X | X |  |  |  |  |  |  |
| Enviar os Resultados Preliminares para a orientadora |  |  |  | X |  |  |  |  |  |  |
| Ajustar os Resultados Preliminares |  |  |  | X |  |  |  |  |  |  |
| Entregar os Resultados Preliminares |  |  |  | X |  |  |  |  |  |  |
| Continuar a analisar os modelos de Machine Learning |  |  |  | X | X | X |  |  |  |  |
| Definir os modelos finais de Machine Learning |  |  |  |  |  | X |  |  |  |  |
| Reunir com a orientadora |  |  |  | X | X | X | X |  |  |  |
| Enviar o Trabalho de Conclusão de Curso para a orientadora |  |  |  |  |  | X | X |  |  |  |
| Ajustar o Trabalho de Conclusão de Curso |  |  |  |  |  |  | X |  |  |  |
| Entregar o Trabalho de Conclusão de Curso |  |  |  |  |  |  | X |  |  |  |
| Preparar a Defesa |  |  |  |  |  |  | X | X |  |  |
| Reunir com a Orientadora |  |  |  |  |  |  | X | X | X |  |
| Enviar a Defesa para a Orientadora |  |  |  |  |  |  |  | X | X |  |
| Ajustar a Defesa |  |  |  |  |  |  |  |  | X |  |
| Entrega e Apresentação da Defesa |  |  |  |  |  |  |  |  | X | X |

Projeto de Pesquisa Resultados Preliminares Entrega do Trabalho de Conclusão de Curso Entrega da Apresentação da Defesa

**Referências**

Axelsson, K.; Gong, J.; Dugast, C.; Lambe, F.; Maquet, P; Suljada, T. 2024. Consumption-based emissions: a new frontier for EU climate policy. SEI Report, June 2024. Stockholm Environment Institute, Stockholm. Disponível em: <https://doi.org/10.51414/sei2024.025>. Acesso em: 1 fev. 2025.

Carbon Cloud. 2019. ClimateHub is an open climate footprint database, that can be used for internal purposes without cost. Disponível em: < https://apps.carboncloud.com/climatehub/>. Acesso em: 1 fev. 2025.

CO2 Everything. 2019. The Carbon Footprint of everyday products and activities. Disponível em: < https://www.co2everything.com/co2e-of/freight-shipping>. Acesso em: 1 fev. 2025.

European Commision. 2013. European Platform on LCA | EPLCA. Disponível em:

<https://eplca.jrc.ec.europa.eu/index.html>. Acesso em 1 fev. 2025.

European Commision. 2021. Consumption Footprint. Disponível em: <https://eplca.jrc.ec.europa.eu/sustainableConsumption.html>. Acesso em 1 fev. 2025.

Favero, L.P.; Belfiore, P. 2024. Manual de Análise de Dados: estatística e Machine Learning com EXCEL®, SPSS®, STATA®, R® e Python®. 2ed. LTC, Rio de Janeiro, RJ, Brasil.

IPCC. 2023. Sections. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 35-115, doi: 10.59327/IPCC/AR6-9789291691647. Disponível em: https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC\_AR6\_SYR\_LongerReport.pdf. Acesso em: 1 fev. 2025.

Malhotra, N. K. 2004. Pesquisa de Marketing: uma aplicação orientada. Bookman, Porto Alegre, RS, Brasil.

Poore, J.; Nemecek T. 2018. Reducing food’s environmental impacts through producers and consumers. Disponível em: <https://www.science.org/doi/10.1126 /science.aaq0216>. Acesso em: 1 fev. 2025.

TBCA. 2013. Tabela Brasileira de Composição de Alimentos. Disponível em: <https://www.tbca.net.br/>. Acesso em: 1 fev. 2025.