Análise Categórica Não Supervisionada de Emissões de Gases de Efeito Estufa e CO_2

Alexandre Cassimiro[1], Diná Xavier[1], Enilda Alves Coelho[1], Kael Soares Augusto[1], Mateus Reis Evangelista[1]

[1] [1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte – MG – Brazil

alexcsilva@hotmail.com, dina.xavier1@gmail.com, enilda.coelho@embrapa.br, kaelsa@hotmail.com.br, 7reis7@gmail.com

Resumo. *Este artigo investiga padrões globais de emissões de CO_2 e gases de efeito estufa a através da abordagem do aprendizado descritivo, utilizando algoritmos de agrupamento e o FP-Growth para extração de regras de associação. Os resultados reforçam que o desenvolvimento econômico está fortemente associado ao aumento das emissões de gases de efeito estufa, ao consumo intensivo de energia e, consequentemente, às mudanças na temperatura global. As associações identificadas entre variáveis econômicas, energéticas e ambientais evidenciam que países com maiores níveis de PIB e industrialização tendem a apresentar maiores emissões e consumo energético, contribuindo de forma que seja significativa para a situação do aquecimento global. As regras de associação extraídas permitiram evidenciar como diferentes fatores interagem e antecedem situações críticas relacionadas ao clima. Dessa forma, o estudo demonstra o potencial do aprendizado descritivo e da mineração de padrões como ferramentas valiosas para a compreensão dos mecanismos que impulsionam as mudanças climáticas, oferecendo subsídios teóricos para a formulação de políticas públicas e estratégias de mitigação ambiental mais eficazes.*

Palavras-chave: Mineração de Padrões, Mudanças Climáticas, FPGrowth, Emissões CO_2.

1. Introdução

O aquecimento global, caracterizado pelo aumento de longo prazo da temperatura média da Terra, representa um dos grandes desafios da atualidade. Existe um consenso, entre diversos pesquisadores, consolidado no relatório do Painel Intergovernamental sobre Mudanças Climáticas (IPCC), que esse fenômeno é inequivocamente causado pelas atividades humanas, principalmente pelas emissões de gases de efeito estufa (GEE), a partir da queima de combustíveis fósseis e industrialização, dentre outros fatores [IPCC 2023]. Esse fenômeno pode trazer graves consequências para a sustentabilidade ambiental, podendo exacerbar ainda mais os eventos climáticos extremos já existentes, como calor extremo, chuvas e secas prolongadas.

A investigação desses fenômenos tem atraído a atenção de pesquisadores do mundo inteiro e de diferentes áreas do conhecimento, que buscam melhor compreensão e mecan-

ismos para minimizar ou eliminar essas ameaças. Nesse contexto de descoberta de conhecimento, a abordagem de aprendizado descritivo, pela aplicação de modelos e algoritmos de mineração de padrões, surgem como uma ferramenta poderosa, potencializada pelo aumento do poder computacional e disponibilidade de grandes bases de dados. Essa abordagem, representa também um grande desafio, com a necessidade de pré-processamento, o alto consumo de recursos computacionais e a interpretabilidade dos resultados, exigindo novas estratégias para modelagem e aprendizado de padrões que sejam contextualmente relevantes e interpretáveis.

Diante desse desafio, novas abordagens de aprendizam descritivo tem se destacado como uma evolução das abordagens clássicas, enfatizando a mineração de padrões diferenciais. O trabalho explora padrões e associações entre variáveis socioeconômicas, energéticas e ambientais, relacionadas à emissão de CO_2, e a relevância de cada padrão, por meio de métricas como confiança, lift e suporte.

Nesse sentido, este trabalho propõe a aplicação da abordagem de aprendizado descritivo, pela aplicação de algoritmos de mineração de padrões e regras de associação no domínio representado pelas emissões de GEE para a descoberta de conhecimento. Com base na abordagem metodológica CRISP-DM [Chapman et al. 2000], foi estruturado em 5 fases que contemplam: entendimento do negócio, entendimento dos dados, preparação dos dados, elaboração de um modelo de aprendizado e avaliação do modelo. O resultado desse trabalho é apresentado nas seções seguintes, que destacam: os estudos relacionados na próxima seção; a abordagem metodológica, na seção 3, os resultados e discussão, na seção 4; e, na seção 5, as considerações finais e sugestão de trabalhos futuros.

Espera-se, portanto, com a demonstração da aplicação dessa abordagem, contribuir para a disseminação e aplicação de algoritmos de aprendizado descritivo promissores para a descoberta de conhecimento.

2. Trabalhos Relacionados

Este trabalho se insere no contexto de aplicações do aprendizado de máquina não supervisionado e mineração de padrões em bases de dados ambientais. Os resultados de revisão de literatura revelam que os estudos se concentram em análises de agrupamentos e, alguns estudos de mineração de padrões frequentes, existindo uma carência de estudos, principalmente estudos que envolvem análises mais aprofundadas em domínios complexos, com múltiplas variáveis, como o domínio dos dados ambientais.

As análises de agrupamentos se destacam com a aplicação do algoritmo K-means. Nesse sentido, [Smith and Zhang 2022] aplica K-means para identificar clusters baseados em emissões per capita e os resultados revelam a dicotomia entre economias industrializadas e emergentes. Por outro lado, os resultados obtidos por [Fathy et al. 2023], [Liu et al. 2023], [Shaban et al. 2024] permitiram correlacionar emissões de carbono com indicadores de desenvolvimento humano por meio de agrupamento hierárquico. Ainda no tema ambiental, [Herman and Shenk 2021] destacou as ferramentas de aprendizado de máquina e descoberta de padrões como uma solução para superar os desafios de trabalhar com a alta dimensionalidade dos dados e complexidades espaciais e temporais de indicadores de políticas ambientais. De um modo geral, os autores ressaltam a importância da mineração de dados e descoberta de padrões e a carência de trabalhos que exploram análises mais complexas.

Para abordagens mais inovadoras e mais aprofundadas, destacam-se novos algoritmos que permitem explorar as bases de dados, não apenas em busca de padrões padrões frequentes, no sentido global, mas também em busca padrões que se diferenciam, caracterizando os subgrupos [Novak et al. 2009]. Uma evolução da descoberta de subgrupos é a mineração de modelos excepcionais (EMM, do inglês Excepcional Modelo Mining). A mineração de modelos excepcionais busca subgrupos onde o modelo que descreve localmente, difere significativamente de um modelo global de referência [Duivesteijn et al. 2016].

Neste sentido, a presente pesquisa se diferencia ao aplicar técnicas de aprendizado descritivo e categorização sobre uma base global e histórica, permitindo a identificação de padrões de coocorrência que revelam não apenas o impacto das emissões, mas também sua relação com o modelo de desenvolvimento econômico de diferentes nações.Ressalta-se que embora tenha sido realizada uma tentativa de aplicação do algoritmo EMM, a principal abordagem adotada foi o algoritmo FP-Growth [Borgelt 2005], buscando aprofundar nas análises. Reconhece-se, contudo, a relevância do EMM e a importância desse trabalho para pavimentar um caminho para trabalhos futuros.

3. Metodologia

Esse trabalho consiste na aplicação do aprendizado descritivo para a descoberta de conhecimento em dados de emissões de CO_2 e gases efeito estufa. As etapas desse trabalho foram realizadas conforme metodologia CRISP-DM [Chapman et al. 2000], ilustrada na Figura 1. Cada etapa é descrita abaixo.

Figure 1. Etapas da metodologia CRISP-DM para Data Science. Fonte: [Chapman et al. 2000]

3.1. Entendendo o negócio

A primeira etapa desse trabalho baseou-se em uma revisão de literatura e um levantamento de fontes de dados, pesquisas em bases de dados de periódicos CAPES e Schoolar Google. Os estudos contemplaram pesquisas fundamentais para a compreensão do tema, para o levantamento dos dados disponíveis e para a seleção dos algoritmos.

Os levantamnentos preliminares evidenciaram a importância do tema, destacando o problema das emissões de CO_2 e gases efeito estufa e sua gravidade e relação com as mudanças climáticas [IPCC 2023].

Um foco foi dado na aplicação de ferramentas de mineração de padrões para análise de dados das emissões e seu impacto dessas emissões no aumento da temperatura. De acordo com o [IPCC 2023], um aumento de 1,5° C na temperatura da terra, pode exacerbar ainda mais, os eventos climáticos extremos, como ondas de calor, secas prolongadas, chuvas torrenciais, inundações, tempestades severas e incêndios florestais, estão se tornando mais frequentes e intensos, exigindo medidas urgentes e efetivas para sua mitigação.

Importante ressaltar que se trata de um tema de interesse global, considerando que a taxa de crescimento de CO_2 na atmosfera (22%) é alarmante e que o aumento da concentração desse gás tem consequente aumento da temperatura afeta. Isso afeta a todos e compromete a sustentabilidade ambiental.

[Friedlingstein et al. 2024], que lidera um Projeto Global de Carbono com 100 pesquisadores de diversos países que monitoram os índices de CO_2, apresenta um total de 40,9 milhões de toneladas de emissões. No âmbito desse projeto, dados de diversos países são organizados e disponibilizados com o objetivo de desenvolver uma imagem completa do ciclo de carbono. Esse conjunto de dados e análises são disponibilizados de forma online e gratuita.

A partir desses dados, a plataforma [Our World in Data 2024] compila novos atributos e disponibiliza, de forma gratuita, com foco nas mudanças climáticas. A base de dados é atualizada regularmente e abrange informações anuais de CO_2, emissões per capita, totais cumulativos e baseados no consumo, conforme ilustra a Tabela 1, destacando a riqueza de atributos e suas características.

Table 1. Relação de atributos do conjunto de dados "CO_2 e gases efeito estufa

Categoria	Atributo	Descrição	Tipo
1. Identificação	`country`, `year`	Identificadores geográfico e temporal da observação.	Categórico, Temporal
2. Econômico/Populacional	`population`, `gdp`	Indicadores demográficos e de Produto Interno Bruto.	Numérico (Contínuo)
3. Emissões por Setor	`coal_co2`, `oil_co2`, `gas_co2`	Emissões anuais de CO_2 por fonte de combustível fóssil.	Numérico (Contínuo)
	`land_use_change_co2`	Emissões de CO_2 devido à Mudança no Uso da Terra (LUC).	Numérico (Contínuo)
4. Emissões Totais	`co2`, `co2_including_luc`	Emissões totais de CO_2, excluindo e incluindo LUC.	Numérico (Contínuo)
5. Crescimento de Emissões	`co2_growth_prct`	Crescimento percentual (ano a ano) nas emissões de CO_2.	Numérico (Percentual)
6. Emissões Per Capita	`co2_per_capita`	Emissões totais de CO_2 por pessoa, para comparação entre países.	Numérico (Contínuo)
7. Intensidade de Carbono	`co2_per_gdp`	Emissões de CO_2 por unidade de PIB.	Numérico (Contínuo)
8. Outros Gases (GHG)	`total_ghg`	Emissões anuais totais de todos os gases de efeito estufa.	Numérico (Contínuo)
9. Emissões Cumulativas	`cumulative_co2`	Emissões históricas totais de CO_2 (responsabilidade histórica).	Numérico (Contínuo)
10. Consumo de Energia	`energy_per_capita`	Consumo de energia primária por pessoa.	Numérico (Contínuo)
11. Participação Global	`share_global_co2`	Participação percentual do país nas emissões globais anuais.	Numérico (Percentual)
12. Mudança de Temperatura	`temperature_change_from_co2`	Impacto estimado das emissões de CO_2 do país na temperatura global.	Numérico (Contínuo)
13. Emissões de Comércio	`trade_co2`	Balanço de emissões de CO_2 do comércio (importações - exportações).	Numérico (Contínuo)

Neste trabalho, optou-se pela utilização da base de dados da plataforma "Our World in Data", pela ampla cobertura e preparação para a abordagem do tema mudanças climáticas.

3.2. Entendendo os dados

A base de dados contempla o período de 1750 a 2022, está em constante atualização e fornece uma base rica em atributos. O dataset possui 79 colunas representando os atributos e 50191 linhas de registros de dados que permitem análises de CO_2 e gases efeito estufa, sob diferentes perspectivas.

O indicador geográfico, presente nesta base, através da coluna ”country” inclui, além de países, referências a agrupamento de localizações geográficas ou geopolíticas, por exemplo, países de renda média alta e continentes.

Embora contemple um longo período de dados, desde a era pré-industrial, o foco desse trabalho foi delimitado a partir de uma análise exploratória e de qualidade dos dados disponíveis. Essas analises contemplaram as estatísticas gerais, assim como também, análises mais detalhadas de cada atributo, avaliando sua completude, distribuição, correlações diversas e análise de relevância para o modelo.

Foi utilizada a linguagem Python, versão 3, para as análises e modelagem. A codificação foi realizada de forma colaborativa e online, utilizando a plataforma Google Colab, um serviço Jupyter Notebook disponível em nuvem, com acesso gratuito a recursos computacionais, incluindo GPUs e TPUs. Um notebook Colab com os scripts, bibliotecas utilizadas e resultados está disponível[1].

A escolha desse ambiente para o desenvolvimento se deu, principalmente, em função da disponibilidade como um serviço em nuvem, de fácil acesso à equipe, com possibilidade de codificação e escrita, de forma colaborativa. Além disso, o Google Colab possui um amplo acesso a bibliotecas de métodos, facilmente incorporados para apoiar todo o pipeline de dados.

3.3. Preparando os dados

As análises exploratórias e de qualidade serviram de orientação para a preparação dos dados. Em uma primeira avaliação da qualidade, foi analisado o percentual de preenchimento de todas as colunas por país.

Para o entendimento dos valores distintos em ”country”, utilizou-se as bibliotecas Geopy e Pycountry para identificar a que tipo de localidade cada valor se refere. Os tipos encontrados foram 123 países, 6 continentes e mais 8 outras classes. Foi avaliado a qualidade dos dados dos tipos diferentes de ’país’. Entendemos que a completude estava baixa e nos concentramos nas localidades tipificadas como ’país’.

Foi realizada a verificação da distribuição de frequências, coluna por coluna. A distribuição da coluna ”year” nos ajudou a ver que a distribuição dos dados não é uniforme para todos os anos. As demais colunas também mostraram distribuição não uniforme.

A análise da completude de todas as colunas a partir de 1965, revelou que melhor qualidade nos dados se concentra entre 1990 e 2020 e que nem todas as colunas têm pelo menos 80% de completude. As figuras 2 e 3 ilustram a completude dos dados, destacando as colunas selecionadas na Figura 3.

[1] https://colab.research.google.com/drive/1BOYlb6UvRNuBOccCWvGicCJymzMmcMEhV?usp=sharing

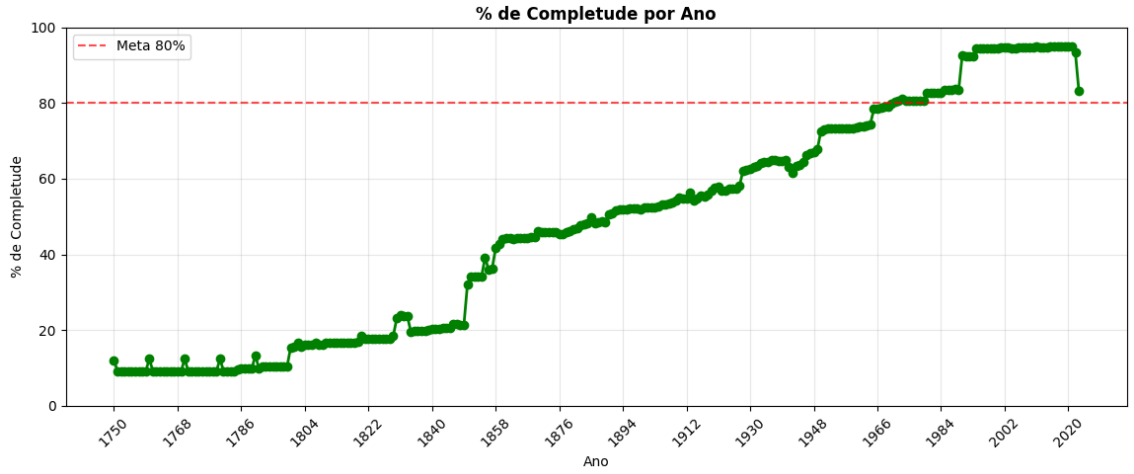

Figure 2. Análise de completude por ano.

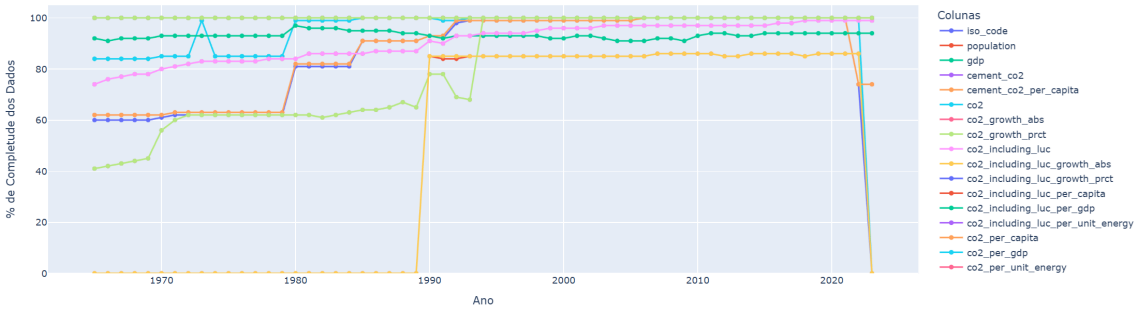

Figure 3. Completude das colunas relevantes a partir do ano de 1965

A análise considerou apenas com dados entre 1990 a 2020 e, para melhoria da qualidade, realizamos remoção de valores nulos e duplicados. Outliers foram visualizados por boxplot. Foi realizada a remoção dos outliers calculando Intervalo Interquartil (IQR) para cada país.

Para utilização do FP-GROWTH, os atributos foram discretizados. Todos os atributos contínuos referente a emissão de gases foram discretizados em 5 bins e 5 labels (muito baixo, baixo, médio, alto e muito alto). Uma amostra do conjunto de dados preparado para análises pode ser visualizada abaixo, na Tabela 2.

Table 2. Amostra do conjunto de dados preparado para análises

	country	year	iso_code	population	population_cat_num	population_cat_label	gdp	gdp_cat_num	gdp_cat_label	cement_co2	...	total_ghg_cat_label	total_ghg_excluding_lucf	total_ghg_excluding_lucf_cat_num
248	Afghanistan	1998	AFG	19159996.0	3	Médio	1.169217e+10	1	Muito baixo	0.047	...	Muito baixo	4.528	1
249	Afghanistan	1999	AFG	19887791.0	3	Médio	1.151732e+10	1	Muito baixo	0.047	...	Muito baixo	4.431	1
250	Afghanistan	2000	AFG	20130334.0	3	Médio	1.128379e+10	1	Muito baixo	0.010	...	Muito baixo	4.521	1
251	Afghanistan	2001	AFG	20284303.0	3	Médio	1.102127e+10	1	Muito baixo	0.007	...	Muito baixo	4.670	1
254	Afghanistan	2004	AFG	23560656.0	3	Médio	2.233257e+10	1	Muito baixo	0.010	...	Muito baixo	5.354	1

A aplicação do algoritmo FP-Growth foi realizada utilizando as principais variáveis relacionadas à temperatura, emissão de gases, consumo de energia (primária e não primária) e economia. Foram selecionadas as seguintes variáveis: `year`, `population`, `gdp`, `co2`, `coal_co2`, `energy_per_capita`, `methane`, `oil_co2`, `total_ghg`, `share_of_temperature_change_from_ghg`.

Foi realizado o cálculo da matriz de correlação entre variáveis numéricas para identificar relações relevantes e em seguinda foi realizada a clusterização dos países, utilizando o algoritmo K-Means, agrupando-os de acordo com as medianas das variáveis numéricas já tratadas e discretizadas.

As variáveis categóricas resultantes da discretização foram transformadas em formato binário (one-hot encoding) utilizando o TransactionEncoder, preparando os dados para a aplicação do algoritmo de mineração de padrões.

3.4. Modelagem

Nesta etapa, após as etapas de limpeza, tratamento, discretização e clusterização dos dados, foi aplicado o FP-Growth [Borgelt 2005], descrito na Tabela 3, com o objetivo de extrair padrões de items frequentes e regras de associação. Os requisitos e pressupostos se mostram adequados às características dos dados obtidos.

Table 3. Características do Algoritmo FP-GROWTH.

Característica	Descrição
Nome Completo	Frequent Pattern Growth
Objetivo	Extrair conjuntos de itens que coocorrem com frequência em um banco de dados transacional, evitando a geração de conjuntos candidatos.
Especificação Formal	Dado um banco de dados de transações D e um limiar de suporte mínimo (minsup), o algoritmo visa encontrar o conjunto de todos os itemsets X tal que o suporte(X) ≥ minsup.
Estrutura de Dados	FP-Tree: Uma estrutura de dados em árvore prefixada que armazena as transações de forma compacta. Nós representam itens e seus contadores, e caminhos compartilhados são mesclados. Tabela de Cabeçalho: Estrutura auxiliar que armazena cada item frequente e um ponteiro para sua primeira ocorrência na FP-Tree.
Pressupostos	• Os dados são transacionais e os itens são categóricos/discretos. • A ordem dos itens dentro de uma transação é irrelevante. • O objetivo é a análise de coocorrência, não de causalidade.
Requisitos	• Dados: Conjunto de dados em formato transacional (lista de listas/conjuntos). • Parâmetros: Definição de um limiar de suporte mínimo (minsup) para a mineração e, opcionalmente, de confiança mínima (minconf) para a geração de regras de associação.
Vantagens	• Desempenho: Geralmente mais rápido que o algoritmo Apriori, pois evita a geração e teste de um número exponencial de candidatos. • Eficiência de Memória: A FP-Tree comprime o dataset, reduzindo a necessidade de memória. • Número de Varreduras: Requer apenas duas varreduras completas no banco de dados original.
Desvantagens	• Complexidade: A estrutura da FP-Tree e o algoritmo de mineração recursivo são mais complexos de implementar e entender. • Uso de Memória em Casos Extremos: Para datasets extremamente esparsos, a FP-Tree pode, paradoxalmente, consumir mais memória. • Reconstrução: Alterar o minsup exige a reconstrução completa da FP-Tree.

Para aplicação do algoritmo e identificação de itemsets frequentes, foi aplicado um suporte mínimo de 10%. Adicionalmente, como métricas para seleção das regras de associação mais relevantes, foi aplicado um filtro, considerando um suporte maior que 15%, Lift maior que 2 e Confiança de 70%. Para cada regra relacionada, foram analisados diversos atributos, incluindo *antecedents*, *consequents* e os resultados de diferentes métricas aplicadas, além do lift, suporte e confiança.

4. Resultados

Na literatura, há um consenso entre diversos pesquisadores sobre a relação entre as emissões antropogênicas de gases de efeito estufa (GEE) e a ocorrência de eventos climáticos extremos, relatado por diversas instituições governamentais e não governamentais, tais como relatam [Friedlingstein et al. 2024] e [IPCC 2023].

Os resultados desse trabalho, com a mineração de padrões para descoberta de conhecimento em dados históricos das emissões de CO_2, além de confirmar os resultados da revisão de literatura, ampliam o olhar para análises mais segmentadas e aprofundadas.

Para a seleção dos padrões mais relevantes, foram adotados os seguintes parâmetros: suporte mínimo de 15%, Lift acima de 2 e Confiança superior a 70%. Como resultado, foram obtidas 210 regras de associação relevantes e não duplicadas.

Para facilitar a interpretação dos resultados, as regras extraídas foram organizadas de acordo com os temas correspondentes às colunas analisadas.

A partir da aplicação dos algoritmos FP-Growth, destacam-se os principais achados conforme o tema do consequente analisado:

Economia: Considerando como consequente, variáveis relacionadas à economia, como o gdp (Gross Domestic Product), observou-se que padrões de altas emissões de CO_2, consumo elevado de energia e altos níveis de gases de efeito estufa frequentemente antecedem situações de economia classificada como "Muito Alto". Ou seja, de forma geral, os resultados indicam que países com economias mais desenvolvidas, representadas por altos valores de PIB (gdp), tendem a apresentar maiores emissões de CO_2, maior consumo de energia e níveis elevados de gases de efeito estufa. Esses resultados estão ilustrados na Figura 4 (Sankey: Antecedentes → Consequente Economia)

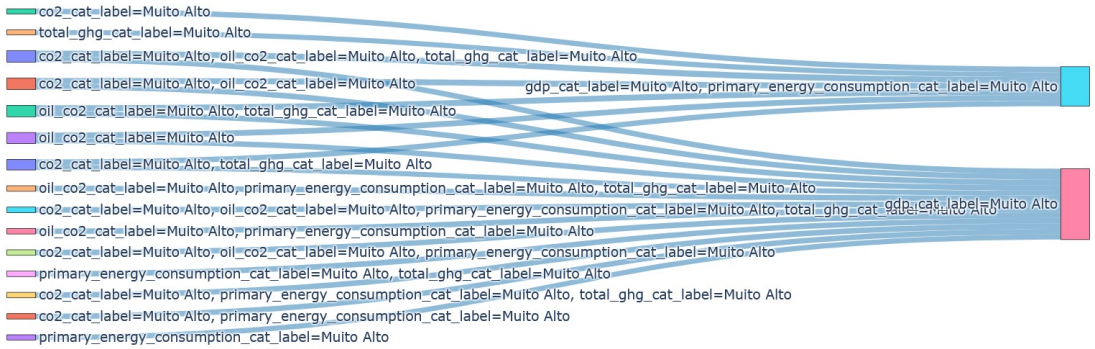

Figure 4. Resultados ilustrados gráfico Sankey: Antecedentes → Consequente Economia.

Energia: Observa-se, nos padrões encontrados no conjunto de itemsets, que combinações como `co2_cat_label=Muito Alto`, `gdp_cat_label=Muito Alto` e `oil_co2_cat_label=Muito Alto`, frequentemente antecedem o consumo primário de energia em níveis "Muito Alto".

Isso indica que países com maior desenvolvimento econômico e maiores emissões de gases de efeito estufa tendem a apresentar também um consumo intensivo de energia primária. Esses resultados reforçam a relação direta entre crescimento econômico, aumento das emissões e demanda energética, indicando que o desenvolvimento econômico, nos moldes atuais, está intrinsecamente ligado ao uso intensivo de recursos energéticos. Estes achados estão representados na Figura 5 (Sankey: Antecedentes → Consequente Energia).

Temperatura: Em relação à temperatura, as regras extraídas mostram que tanto

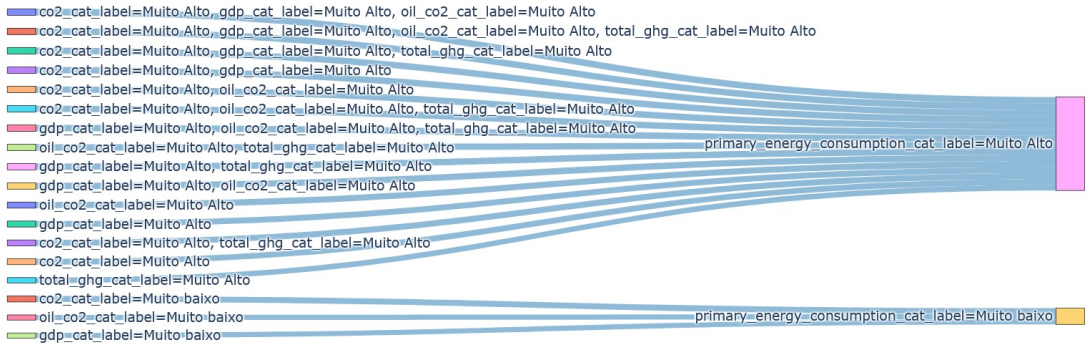

Figure 5. Resultados ilustrados gráfico Sankey: Antecedentes → Consequente Energia.

baixos quanto altos níveis de gases de efeito estufa (GHG) e metano estão associados a variações de temperatura. Por exemplo, observa-se que emissões elevadas de CO_2 e GHG frequentemente antecedem aumentos significativos na participação das mudanças de temperatura atribuídas aos gases de efeito estufa. Os principais padrões estão apresentados na Figura 6 (Sankey: Antecedentes → Consequente Temperatura).

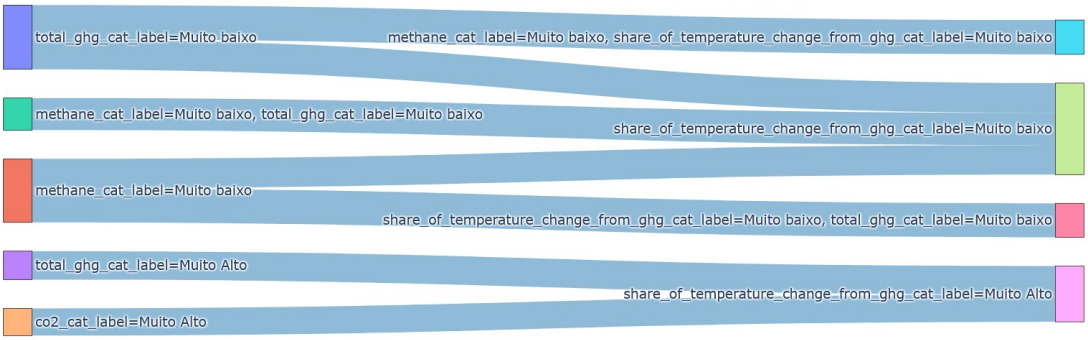

Figure 6. Resultados ilustrados gráfico Sankey: Antecedentes → Consequente Temperatura.

Adicionalmente, buscou-se identificar, ao longo de todo o período avaliado, se haveria frequência de anos em algum país com comportamento inesperado, porém, de acordo com as métricas de avaliação adotadas, nenhum resultado relevante foi encontrado para esse critério.

4.1. Discussão dos Resultados

Os resultados obtidos reforçaram o conceito de que desenvolvimento econômico está fortemente associado ao aumento das emissões de gases de efeito estufa, ao consumo intensivo de energia e, consequentemente, às mudanças na temperatura global. As associações identificadas entre variáveis econômicas, energéticas e ambientais evidenciam que países com maiores níveis de PIB e industrialização tendem a apresentar maiores emissões e consumo energético, contribuindo de forma significativa para o aquecimento global. As regras de associação a partir desta aplicação, permitiram evidenciar como diferentes fatores interagem e antecedem situações críticas relacionadas ao clima. Além disso, a aplicação

das técnicas de clusterização e mineração de padrões mostrou-se de grande utilidade para revelação tanto de padrões globais quanto de associações específicas, que seriam dificilmente identificadas por meio de análises univariadas ou descritivas. A utilização desta abordagem integrada prmitiu a exploração das inter-relações entre variáveis, de forma a ampliar o potencial explicativo do estudo. Dessa forma, conclui-se que este estudo demonstra o potencial do aprendizado descritivo e da mineração de padrões como ferramentas valiosas para a compreensão dos mecanismos que impulsionam as mudanças climáticas. A partir da aplicação dessas técnicas em grandes bases de dados ambientais, acredita-se ser possível obter insumos teóricos que contribuam para a formulação de políticas públicas e estratégias de mitigação de danos ao meio ambiente mais eficazes, auxiliando no enfrentamento dos desafios impostos pelo aquecimento global.

4.2. Desafios e limitações

A aplicação do algoritmo FP-Growth apresentou-se adequada ao objetivo de descoberta de padrões proposto. No entanto, alguns desafios foram identificados ao longo do processo, principalmente relacionados à organização das informações, devido ao grande volume e à variabilidade dos dados e à interpretação dos padrões e regras extraídos. Como o FP-Growth opera exclusivamente com dados categóricos, foi imprescindível a realização da etapa de normalização e discretização, classificando as variáveis em categorias como ”muito baixo”, ”baixo”, ”médio”, ”alto” e ”muito alto”. Essa transformação foi fundamental para a identificação dos padrões e para a interpretação dos resultados. Porém, dentre as limitações observadas, destaca-se a possibilidade de que a discretização dos dados possa ter levado à perda de nuances importantes, dificultando a identificação de padrões mais sutis ou inesperados. Outro aspecto relevante, refere-se à presença de dados incompletos em diversas situações, que precisaram ser tratados e limpos no dataset, o que pode ter limitado a identificação de padrões inesperados. Alternativamente, a ausência desses padrões pode indicar que, de fato, tais comportamentos não existem de forma significativa nos dados analisados.

5. Considerações finais

Este trabalho apresentou uma metodologia de mineração de dados e os resultados de uma abordagem de aprendizado descritivo, pela mineração de padrões em dados globais de emissões de CO_2 e gases de efeito estufa. Como contribuição principal, destaca-se a aplicação do aprendizado descritivo, demonstrado através de algoritmos de mineração de padrões, o FP-Growth, e o grande potencial da ferramenta para a descoberta de conhecimento relevante, específico e útil.

As análises realizadas demonstraram o a eficácia dessa abordagem na revelação de padrões significativos. A discretização das variáveis contínuas permitiu representar de forma categórica e comparável as diferentes dimensões socioeconômicas e ambientais, favorecendo as análises e interpretação dos resultados. A partir da aplicação do FP-Growth, os padrões identificados evidenciaram relações relevantes, além de revelar nuances importantes, como correlações negativas localizadas, que sugerem interações complexas e, por vezes, contraintuitivas no cenário climático global. Esses achados corroboram evidências já estabelecidas na literatura sobre o impacto das atividades humanas no aquecimento global. Complementando os achados apresentados, a visualização de correlações positivas e negativas entre variáveis ambientais e socioeconômicas reforça a importância de se

considerar múltiplos fatores no entendimento da dinâmica climática. Assim, este trabalho contribui para o avanços na análise descritiva de dados ambientais e para construção de um debate mais embasado e multidimensional sobre as mudanças climáticas.

Do ponto de vista técnico, a aplicação das técnicas de clusterização e, especialmente, do algoritmo FP-Growth, devido à sua grande utilidade na revelação de padrões, demonstrou-se uma ferramenta de grande eficácia e adequado a este tipo de extração de regras e padrões interpretáveis. Além disso, este trabalho contribui para o reconhecimento do modelo CRISP-DM como um processo orientador para a mineração de dados. Dessa forma, reforça a importância de empregar técnicas de mineração de padrões em grandes bases históricas de dados ambientais, possibilitando uma compreensão mais aprofundada e contextualizada das emissões.

Reconhecemos, entretanto, que apesar de todos os resultados satisfatórios obtidos a partir da aplicação do FP-Growth no presente estudo, tornou-se evidente uma série de desafios e limitações, como é o caso da discretização das variáveis para aplicações do algoritmo, que pode ter levado a uma perda de nuances importantes, presentes nos dados. A variabilidade das informações exigiu cuidados adicionais na realização de análise e tratamento dos dados, para que fosse evitado uma limitação na identificação dos padrões inesperados.

Como trabalhos futuros, o objetivo principal é aprofundar na melhoria da qualidade dos dados, na estruturação e aplicação da Mineração de Modelos Excepcionais (EMM) em busca de novos insights.

References

[Borgelt 2005] Borgelt, C. (2005). An implementation of the fp-growth algorithm. Technical report, Otto-von-Guericke-University of Magdeburg, Department of Knowledge Processing and Language Engineering, Magdeburg, Germany.

[Chapman et al. 2000] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. [S. l.].

[Duivesteijn et al. 2016] Duivesteijn, W., Feelders, A., and Knobbe, A. (2016). Exceptional model mining. *Data Mining and Knowledge Discovery*, 30:47–98.

[Fathy et al. 2023] Fathy, A., Elaziz, M. A., Zhang, P., Elkenawy, E.-S. M., Hassanien, A. E., Faris, H., and Mirjalili, S. (2023). A novel blockchain-based architecture for smart grid energy trading using deep reinforcement learning. *Journal of Ambient Intelligence and Humanized Computing*, 15:2969–2983.

[Friedlingstein et al. 2024] Friedlingstein, P. et al. (2024). Global carbon budget 2024. *Earth System Science Data*, 16:2549–2605.

[Herman and Shenk 2021] Herman, K. S. and Shenk, J. (2021). Pattern discovery for climate and environmental policy indicators. *Environmental Science Policy*, 120:89–98.

[IPCC 2023] IPCC (2023). Summary for policymakers. 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*, pages 1–34. IPCC, Geneva, Switzerland.

[Liu et al. 2023] Liu, Y., Zhang, Y., Liu, L., Fu, H., Yu, H., and Li, J. (2023). Carbon emissions and driving forces in china's urban agglomerations: New insights from multisource data. *Journal of Environmental Sciences*, 132:264–278.

[Novak et al. 2009] Novak, P. K., Lavrač, N., and Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403.

[Our World in Data 2024] Our World in Data (2024). OWID CO_2 and Greenhouse Gas Emissions Dataset [conjunto de dados]. [S. l.]: Our World in Data.

[Shaban et al. 2024] Shaban, M. A., El-Moselhy, A. M., and Gouda, M. M. (2024). Analysis and prediction of carbon dioxide emissions based on socioeconomic indicators using machine learning techniques. *Journal of Civil Engineering and Sustainable Environment*, 7(1):43–58.

[Smith and Zhang 2022] Smith, J. and Zhang, L. (2022). Clustering national co_2 trajectories: A k-means approach. *Environmental Science & Policy*, 135:45–53.