Análise categórica não supervisionada de gases efeito estufa.

- Our World in Data

Alexandre Cassimiro; Diná Xavier; Enilda Alves Coelho; Kael Soares Augusto; Mateus Reis Evangelista;

Roteiro

Entendendo os dados

Análise e preparação

Algoritmo escolhido

Resultados

Considerações finais

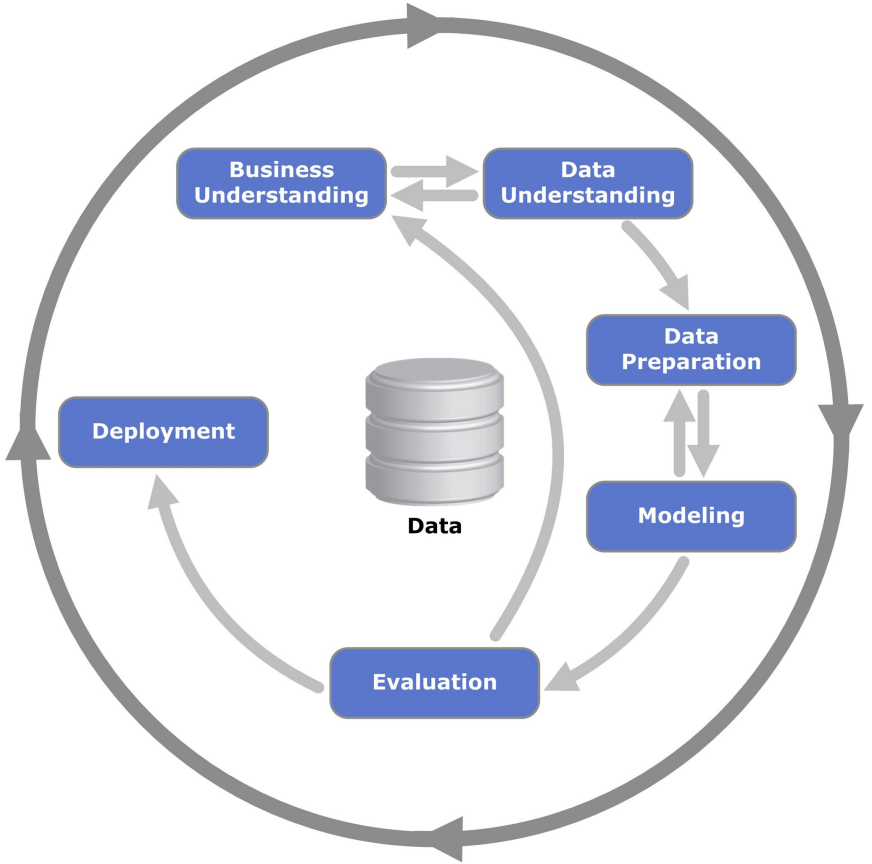


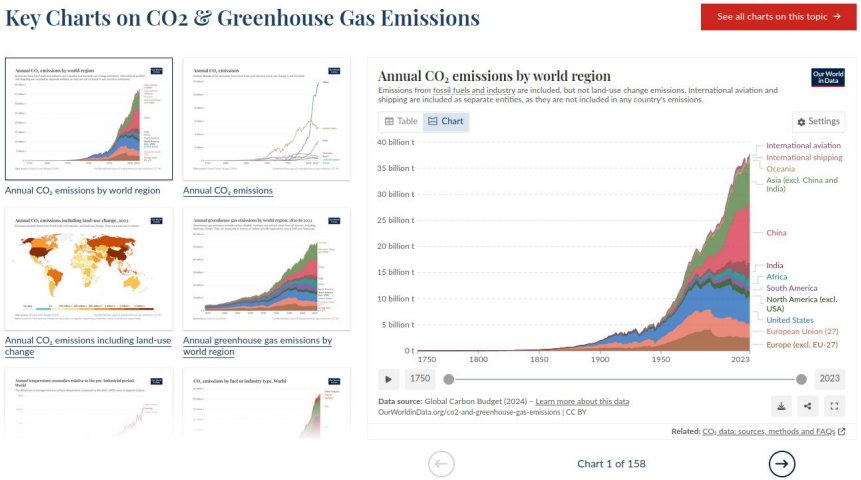
Figura 1. Etapas da metodologia CRISP-DM para Data Science. Fonte: Chapman et al. (2000)

Entendendo os dados:

CO_2 and Greenhouse Gas Emissions

Our World in Data:
https://ourworldindata.org/co2-and-greenhouse-gas-emissions

Os dados foram adquiridos da fonte Our World in Data:



Note os 158 gráficos feitos com os dados!

Formato dos dados - Entrada Original

No total uma matriz de 50K~ (baseados em <Localização x Ano>) linhas e 79 (atributos) colunas.

Dentre os atributos, temos notoriamente:

Quadro 1. Relação de atributos do conjunto de dados "CO² 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)

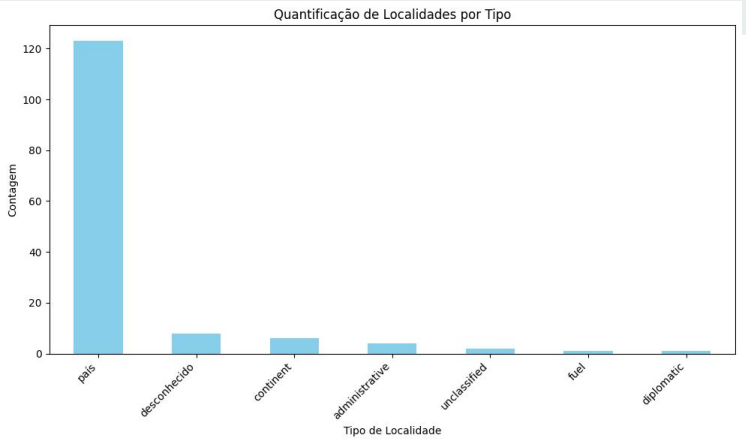
e muitas variações:

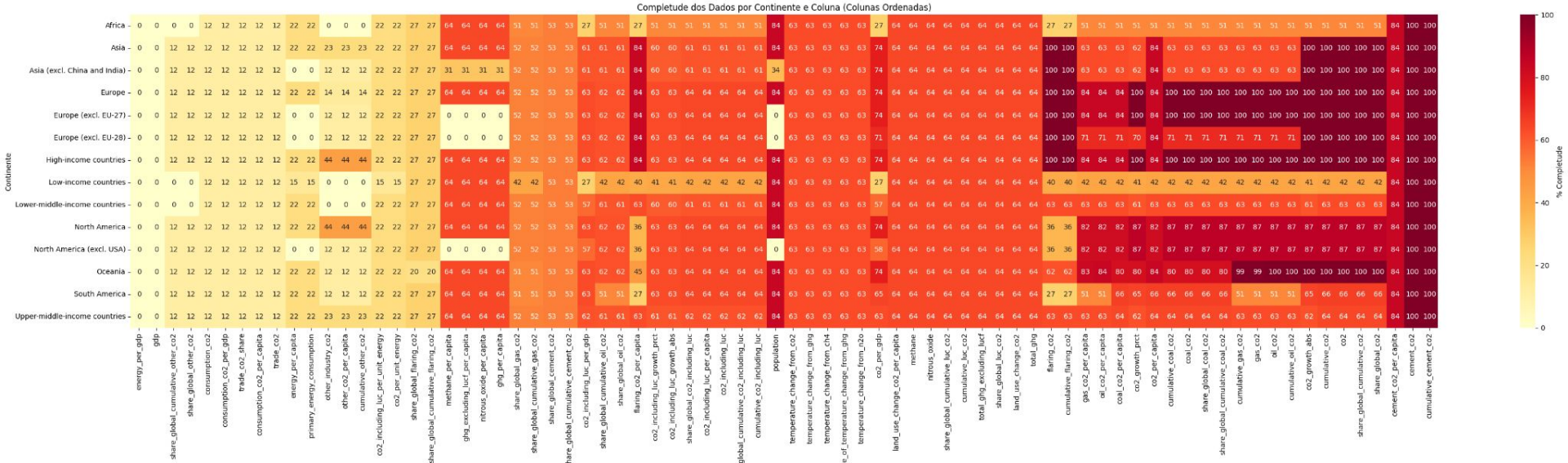
```
co2
co2_growth_abs
co2_growth_prct
co2_including_luc
co2_including_luc_growth_abs
co2_including_luc_growth_prct
co2_including_luc_per_capita
co2_including_luc_per_gdp
co2_including_luc_per_unit_energy
co2_per_capita
co2_per_gdp
co2_per_unit_energy
```

Dados estranhos e nulos:

Algumas localidades dos dados não foram países.

Além disso, a completude dos dados, em especial para continentes, não era muito boa:





Análise e preparação de dados:

Limpeza e categorização

Limpeza de dados nulos e categorização dos dados.

Os dados entregues têm bastante espaço vazio. Para ter uma análise adequada, se torna necessário fazer transformações que resultem em dados.

Além disso, devido ao algoritmo que escolhemos, os dados precisam de ser categóricos, necessitando uma transformação dos dados originalmente inteiros.

Limpeza de dados: Cobertura de Dados x Ano

As primeiras décadas do dataset são **muito pouco** preenchidos para a grande maioria dos atributos. Decidimos apenas analisar os dados a partir da década de 90 em diante.

Favor notar que no gráfico muitos atributos estão sobrepostos entre si.

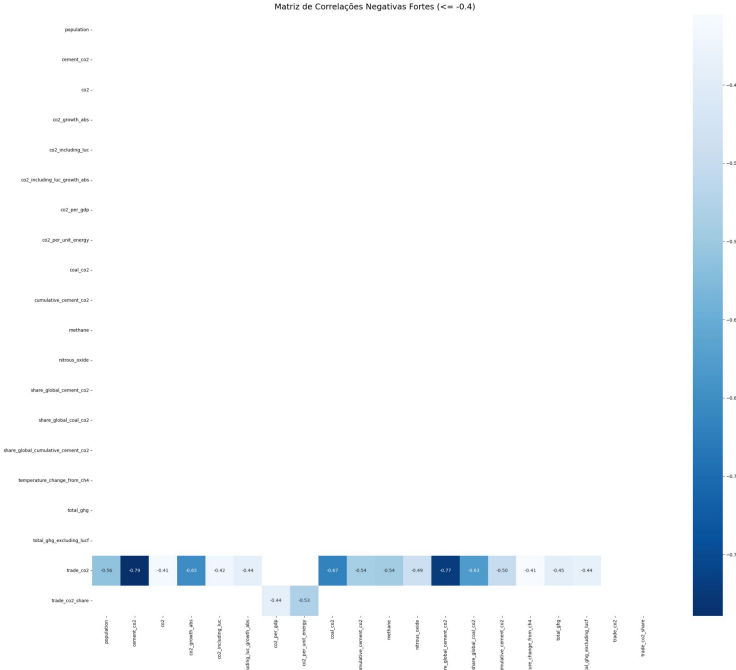
Remoção de Outliers e Categorização dos Dados

Foi usado a técnica do `Intervalo Interquartil (IQR) <Q1 - 1.5*IQR> & <Q3 + 1.5*IQR>` para remover outliers fortes do banco de dados.

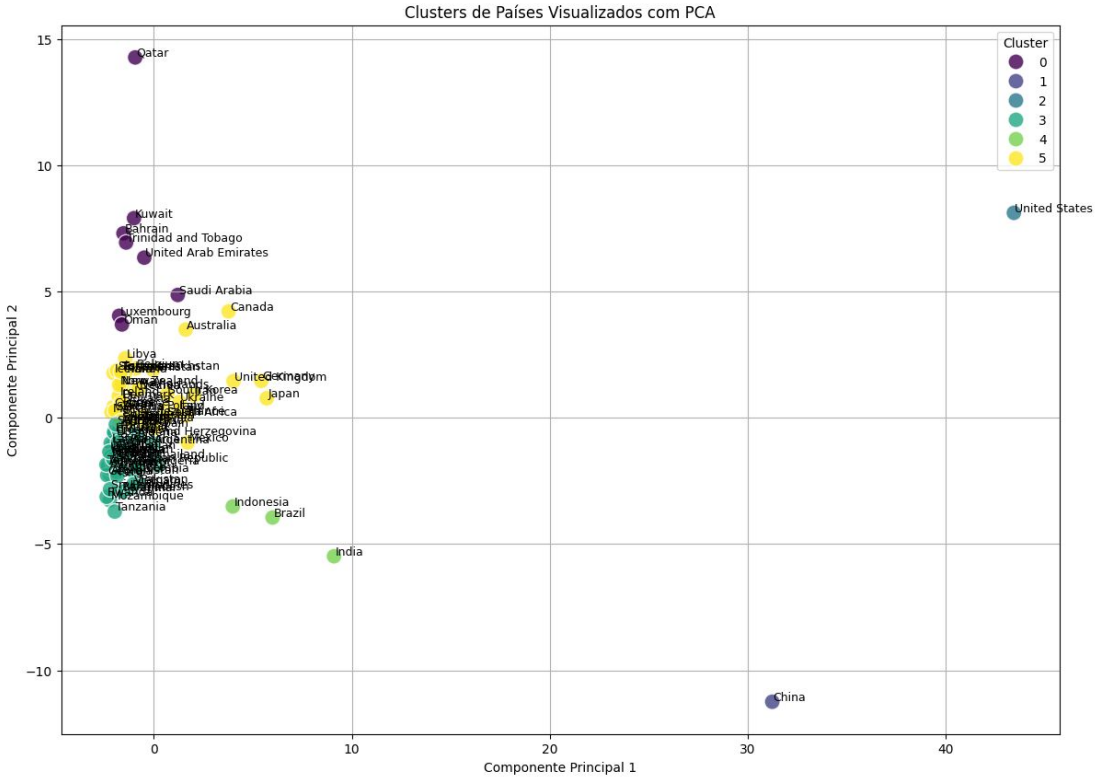
Os dados restantes foram categorizados usando os 5 quartis entre:

- **Muito Baixo**
- **Baixo**
- **Médio**
- **Alto**
- **Muito Alto**

Correlações de Dados



Clusterização dos Dados



Algoritmo Executado

FP-Growth

Aprendizado não supervisionado de gases efeito estufa

O algoritmo escolhido para analisar os dados foi o FP-Growth. Outros algoritmos como por exemplo o EMM foram testados, mas sem resultados interessantes.

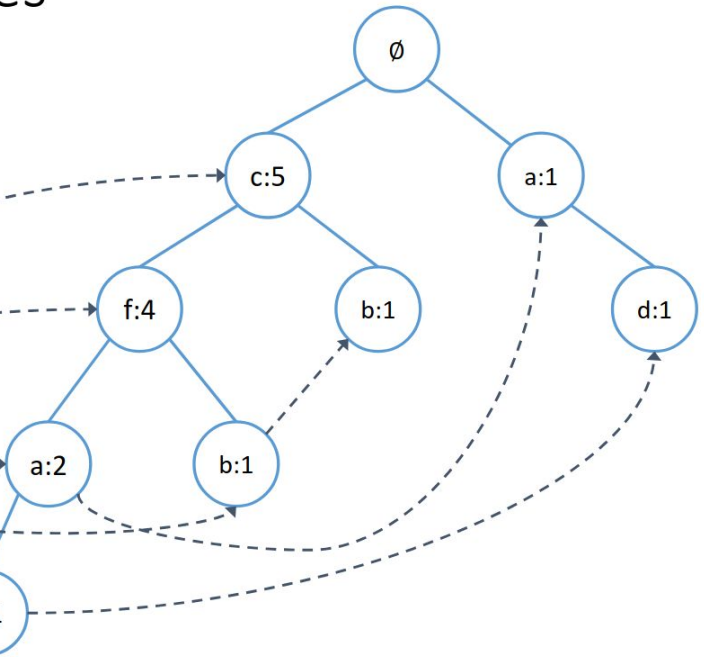
Contudo, o FP-Growth obteve resultados! Primeiro vamos para o uso do algoritmo e depois a análise.

FP-Growth - Lembrete

Mineração dos padrões

- Exemplo

Item	Freq	Link
c	5	
f	4	
a	3	
b	2	
d	2	



FP-Growth - Implementação

```
from mlxtend.frequent_patterns import fpgrowth
```

Usamos a implementação do mlxtend.frequent_patterns para o FP-Growth. Isso nos entrega o conjunto de itemsets frequentes que temos no algoritmo. Além disso, usamos o

```
from mlxtend.frequent_patterns import association_rules
```

para gerar as regras a partir do resultado do FP-Growth, assim chegando nas regras de associação, que são o nosso objetivo final!

Por final, ordenamos o resultado pelas métricas de avaliação para chegar no resultado final.

Resultados Finais

Resumo dos resultados principais do algoritmo em relação com o esperado.

Antes de analisar os dados, já haviam várias expectativas dos resultados que teríamos. Coisas como por exemplo aumento de gases de efeito estufa levando à maior temperatura. Países com menor GDP afetam menos o efeito estufa quando comparado com o resto, etc.

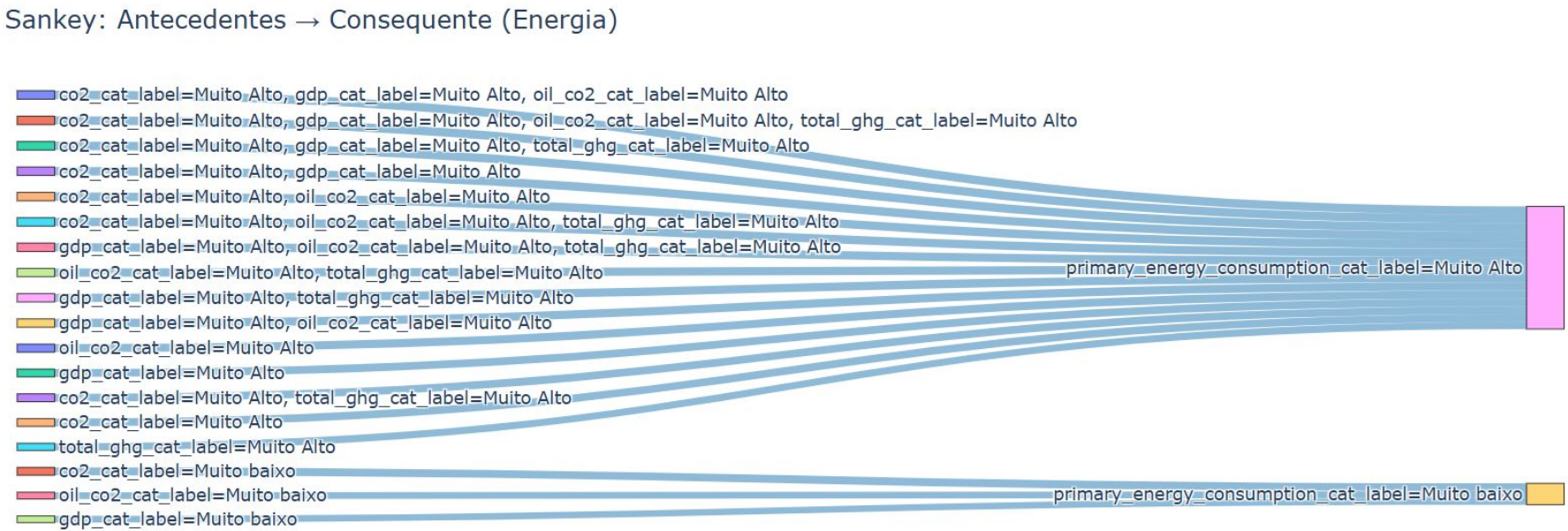
Uma das coisas interessantes dessa análise é poder validar se os conceitos que já temos se encontram nos dados e procurar análises que talvez sejam diferentes.

No final também colocamos a tentativa do EMM.

Resultados Finais

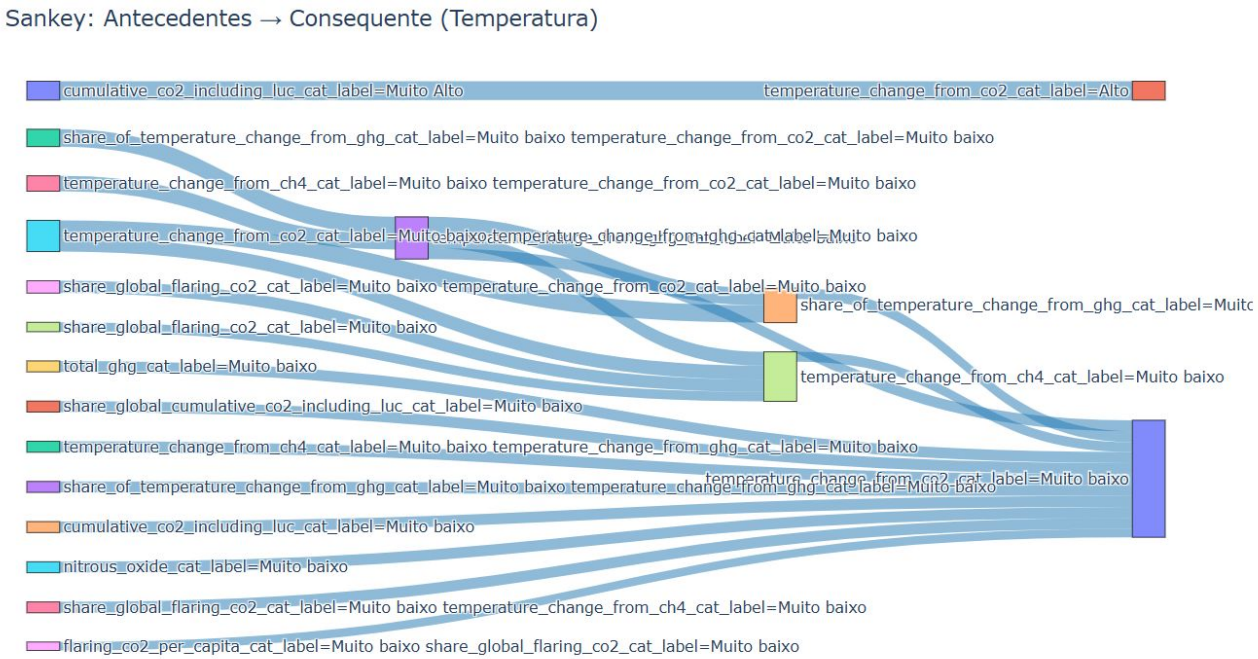
Métricas de avaliação: Support > 0.15 Lift> 2 Confidence > 0.7

Resultados relacionados à Energia Suporte fp-growth 10%

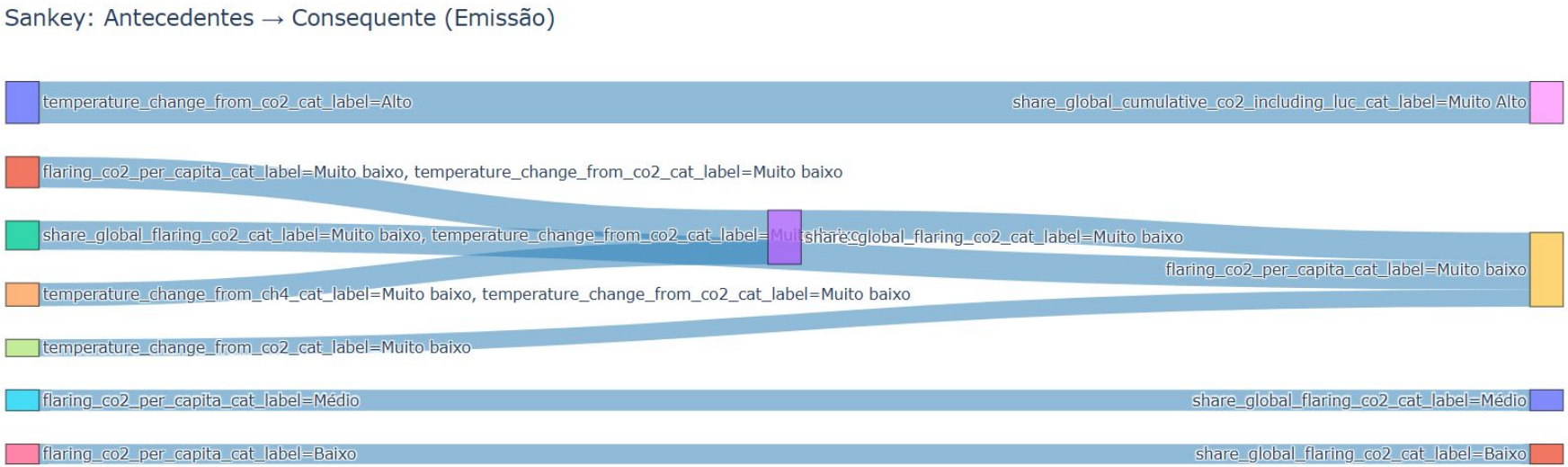


Resultados Finais

Resultados ligados à mudança de temperatura



Resultados relacionados à Emissão de gases

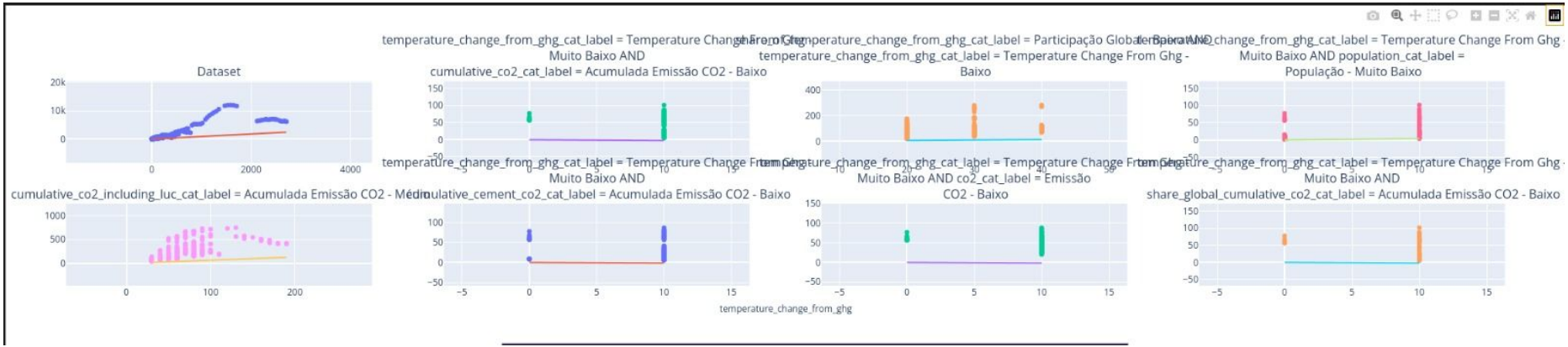


Resultados Finais

Emissão de “Outros gases”

EMM - Tentativa Falha

Apesar da tentativa de usar Exceptional Model Mining, não obtivemos resultados bons. Como um exemplo, considere o seguinte gráfico obtido:



Conclusões

Takeaways principais do trabalho, overview geral e entrega final.

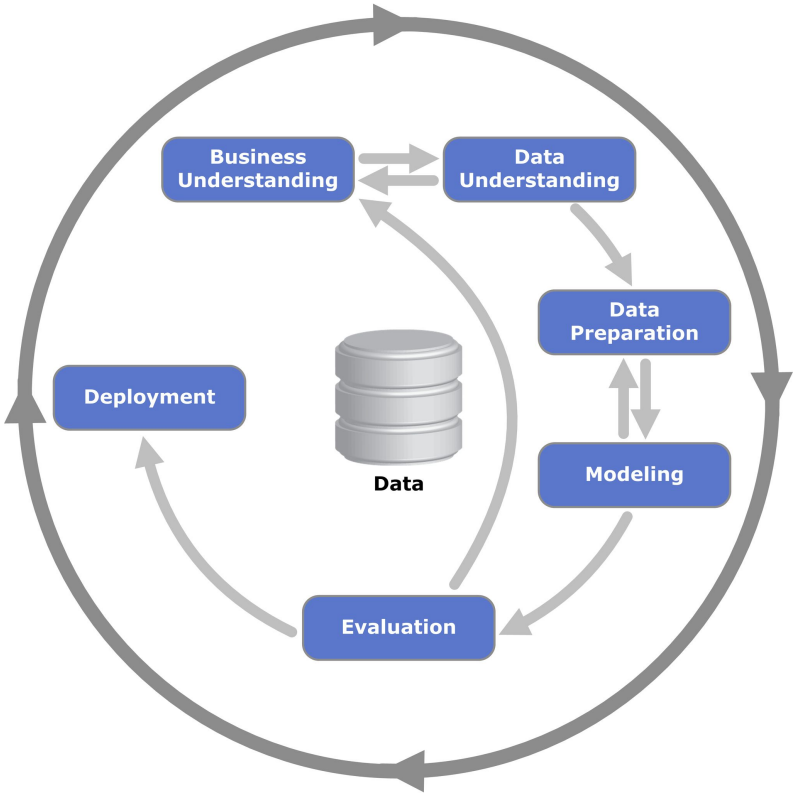
Ao longo do trabalho muitas coisas foram feitas. Algumas deram certo, outras não. Essa seção é dedicada às expectativas, divergências do original, ideias implementadas e uma revisão geral de tudo que foi dito.

Metodologia

Esse trabalho foi uma oportunidade para o grupo aprender a fazer análise de dados em conjunto, coisa que precisa de comunicação, colaboração e uma metodologia clara para ser eficiente.

Ao longo desse processo, o grupo obteve vários insights de como formar análises e definir estruturas e modelos adequados.

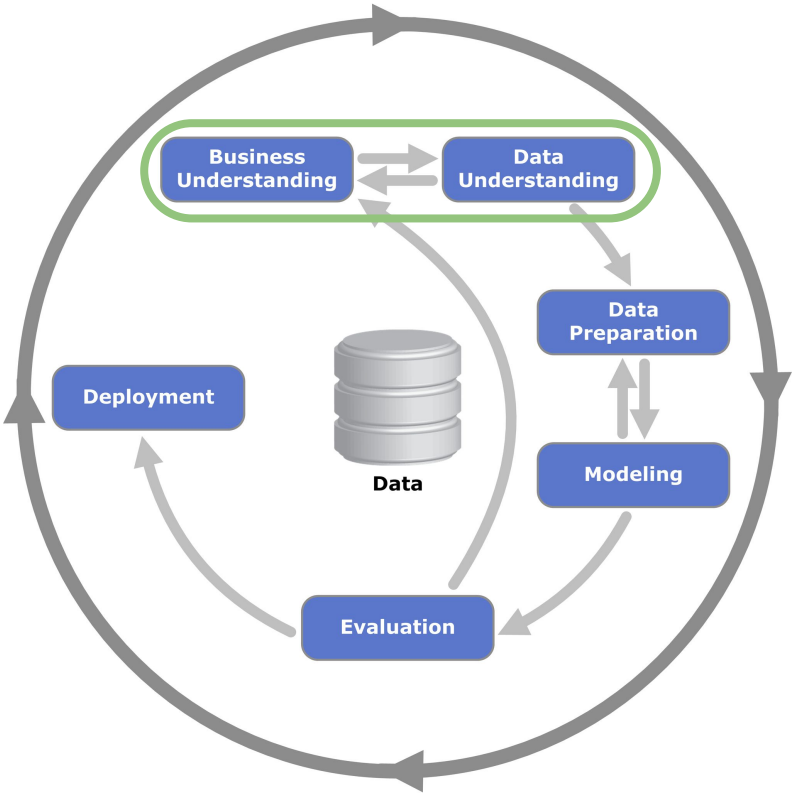
A metodologia escolhida foi a Crisp-DM, que foi usada a partir da fase de escolhida de dados. Podemos ver à direita os passos da Crisp-DM. Vamos recontar eles um a um agora da perspectiva do grupo.



Business/Data Understanding

Os dados foram escolhidos devido à sua relevância social: O efeito estufa é uma preocupação global crescente. O nosso objetivo é o de fazer análises que a mineração de dados, como vista no curso, disponibiliza. Mais especificamente, objetivamos encontrar análises não antes vistas.

No nosso caso, temos dados do tipo de inteiros ou floats descrevendo vários atributos, países como categorias e anos.



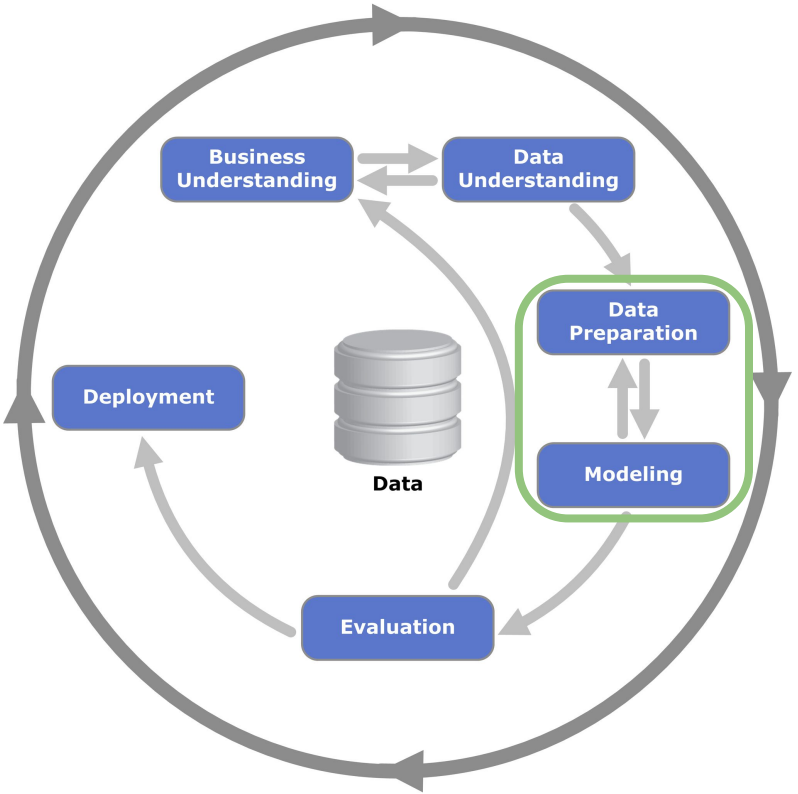
Data Preparation/Modeling

A escolha do modelo não foi trivial. Boa parte das escolhas tinham trade-offs. Escolhemos preparar os dados para dois modelos em específico:

FP-Growth - Usa dados categóricos, não supervisionado. Encontra regras.

EMM - Usa dados categóricos ou discretizados, supervisionado com variáveis alvos.

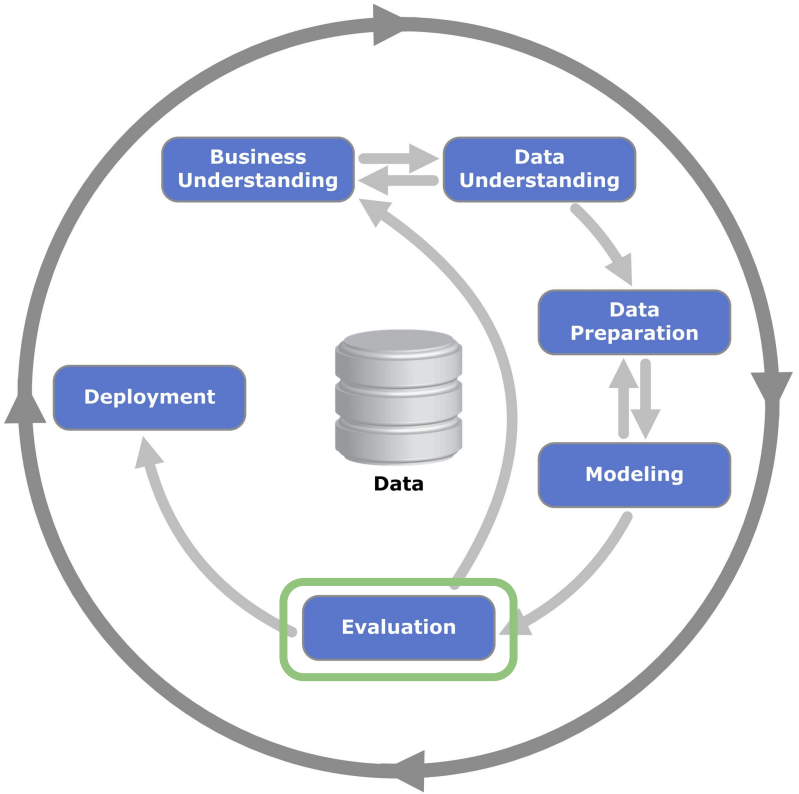
Como ambos usam ou podem usar dados categóricos, decidimos adaptar os dados para este cenário.



Evaluation

Os dados foram avaliados e plotados. No caso do EMM, retirado do https://github.com/MathynS/emm, não obtivemos resultados interessantes. Suspeitamos que isso seja devido ao formato dos dados e a interligação forte entre várias das colunas.

No caso do FP-Growth, obtivemos um número bom de regras! A partir disso plotamos gráficos Sankey para a visualização dos resultados, que foi apresentada acima.



Evaluation - Imprevistos

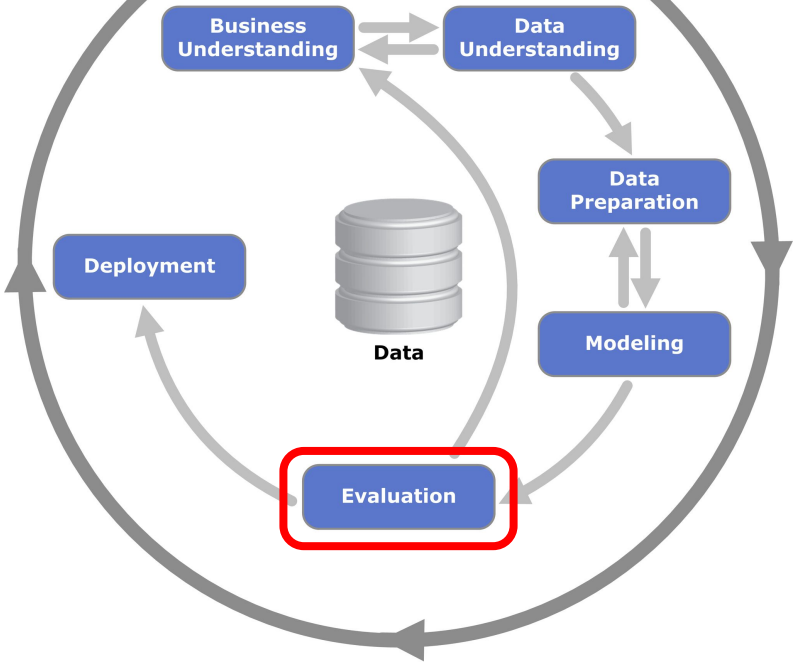
Temos 79 colunas. Das quais, muitas são uma versão cumulativa da outra, ou uma razão entre duas, etc.

De início isso gerava resultados muito triviais, como growth_abs -> growth_prct. Para resolver este problema decidimos reduzir o número de colunas analisadas para as principais, evitando as interligadas:

```
'population' 'gdp' 'co2' 'coal_co2'
'energy_per_capita' 'methane' 'oil_co2'
'primary_energy_consumption' 'total_ghg'
'share_of_temperature_change_from_ghg'
```

Para algumas das análises.

```
co2
co2_growth_abs
co2_growth_prct
co2_including_luc
co2_including_luc_growth_abs
co2_including_luc_growth_prct
co2_including_luc_per_capita
co2_including_luc_per_gdp
co2_including_luc_per_unit_energy
co2_per_capita
co2_per_gdp
co2_per_unit_energy
```

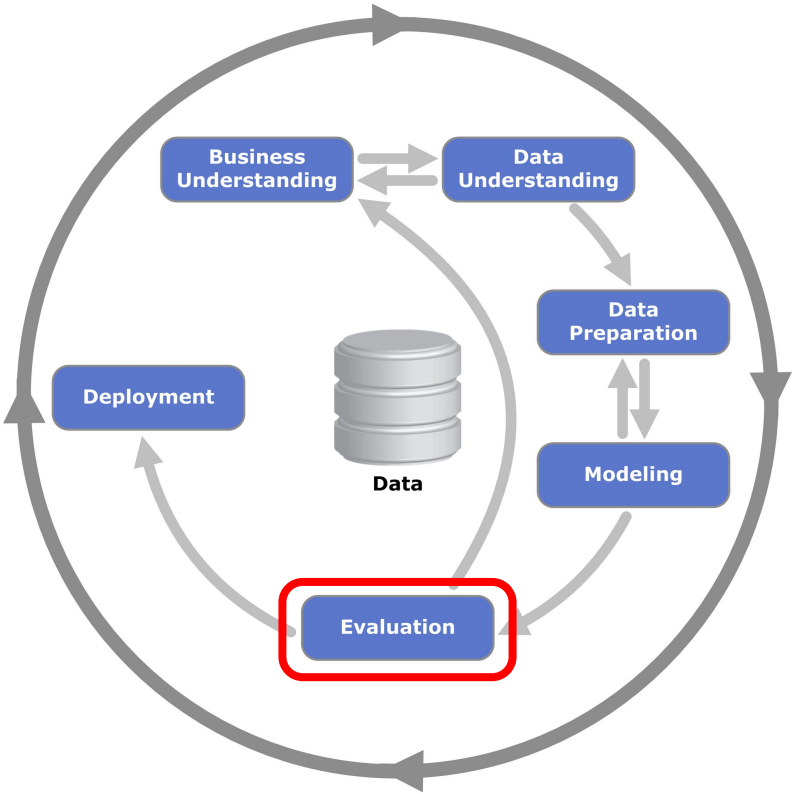


Evaluation - Críticas

Grande parte dos resultados que o FP-Growth retornou para a nossa análise foram bem simples: Regras de 1:1, ou de poucos para poucos.

Suspeitamos que isso pode ser em parte devido ao fato que o FP-Growth não lida muito bem com dados forçadamente categorizado por meio de quartis.

Outro motivo pode ser apenas uma característica dos dados: ele não possui regras complexas com suporte alto o suficiente para existir.



Evaluation - Escolha Final

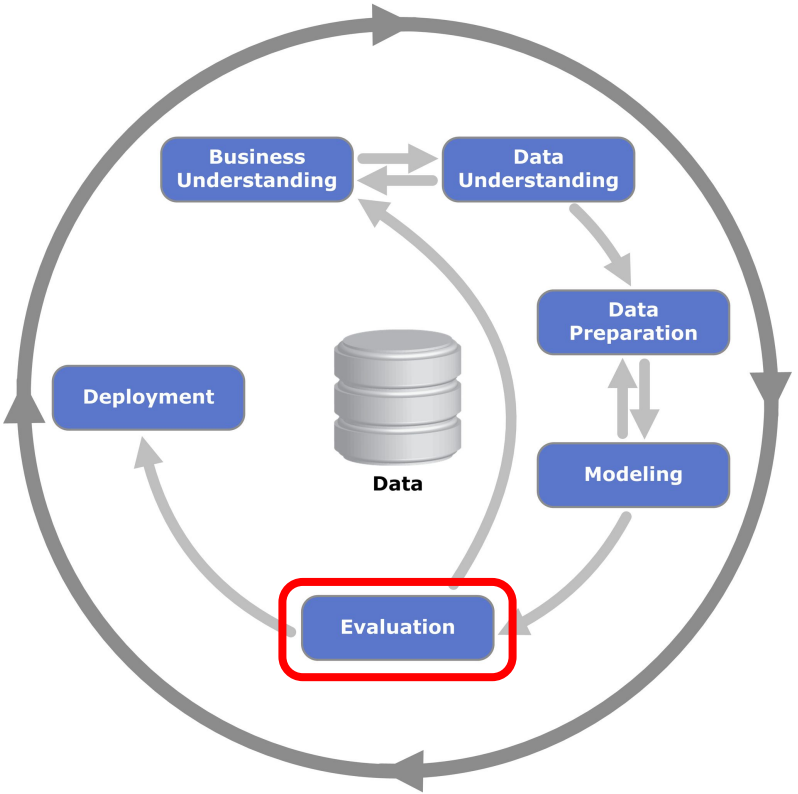
Apesar das desvantagens, o FP-Growth ainda foi um dos modelos que melhor se adequou para a situação que tentamos abordar. Em sumo:

Vantagens:

- Simples implementação/uso
- Resultados plotáveis
- Métricas reguláveis

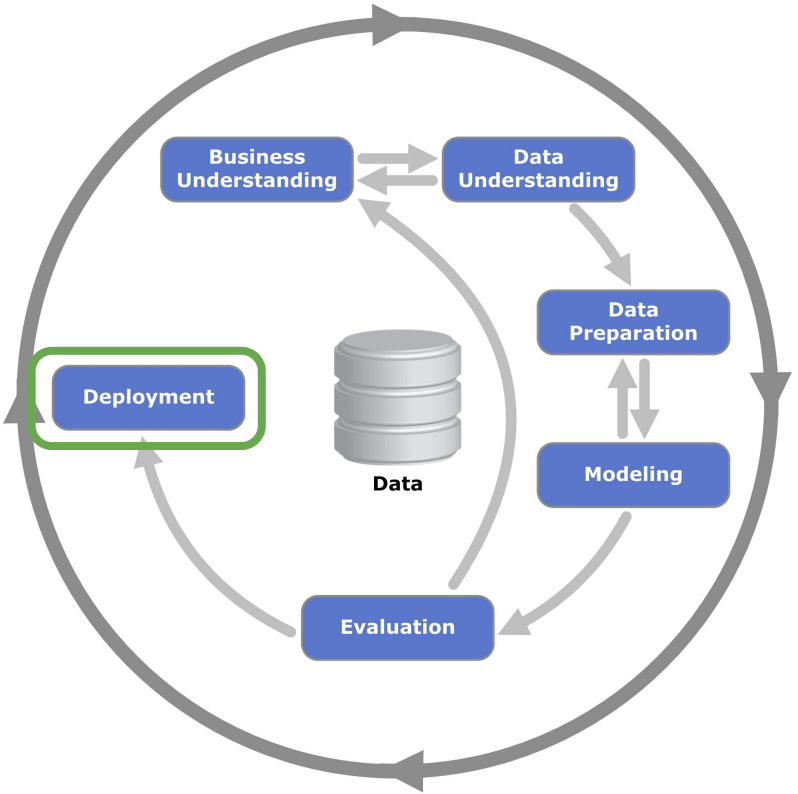
Desvantagem

- Não lida bem com dados categorizados em buckets de quartis
- Resultados simples (nesse caso)
- Muitos resultados similares



Deployment

O notebook com o código e o notebook que inspirou o artigo da entrega se encontram em https://github.com/Dwctor/AD_Analise_CO2. No repositório haverá dois notebooks (código e artigo), o artigo final, essa apresentação e um arquivo requirements.txt com todas dependências e suas versões em python para executar os notebooks, tornando ambos reproduzíveis.



Obrigado pela atenção!!!

Dúvidas, comentários ou sugestões?