Sugerindo colaborações na base de doutores da Plataforma Lattes usando Florestas Aleatórias

Gabriel Dahia, Gabriel Lecomte, Pedro Vidal

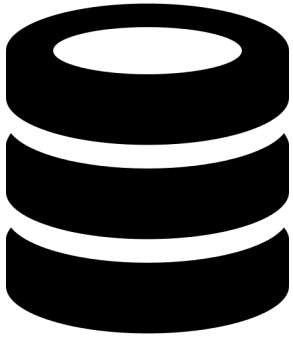
Seminário - MATA60 - UFBA

Introdução

Plataforma Lattes

LattesDoctoralDataset

- 265.187 doutores;
- Número e tipo de publicações;
- Colaborações;
- Atuação Profissional;
- Formação acadêmica.



Objetivo

"Desenvolver um método de mineração de dados para sugerir ou prever novas colaborações com base nos dados existentes"

Colaborações atualmente: **apenas 0,02% do total possível**.

Metodologia

Pré-processamento

Descarte de currículos

- Ausência de colaborações;
- Valores inválidos:
  - Atuação profissional;
  - Formação acadêmica.
- Ausência de produção científica;
- Campos com significado incerto.

Campos de preenchimento livre

“USP”

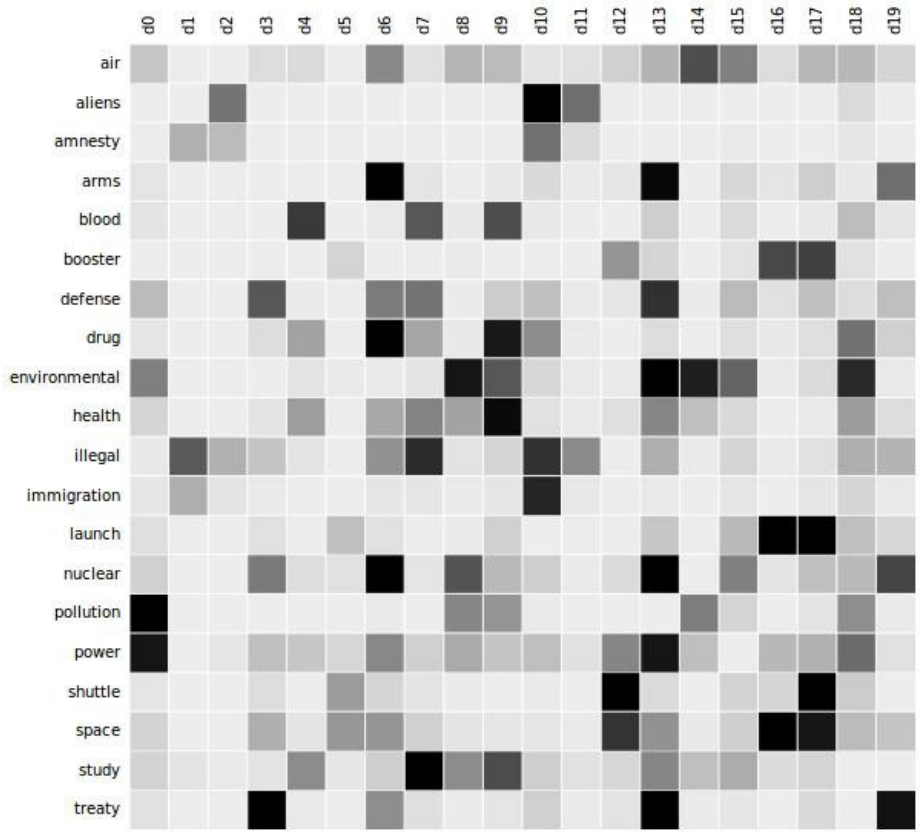
ou

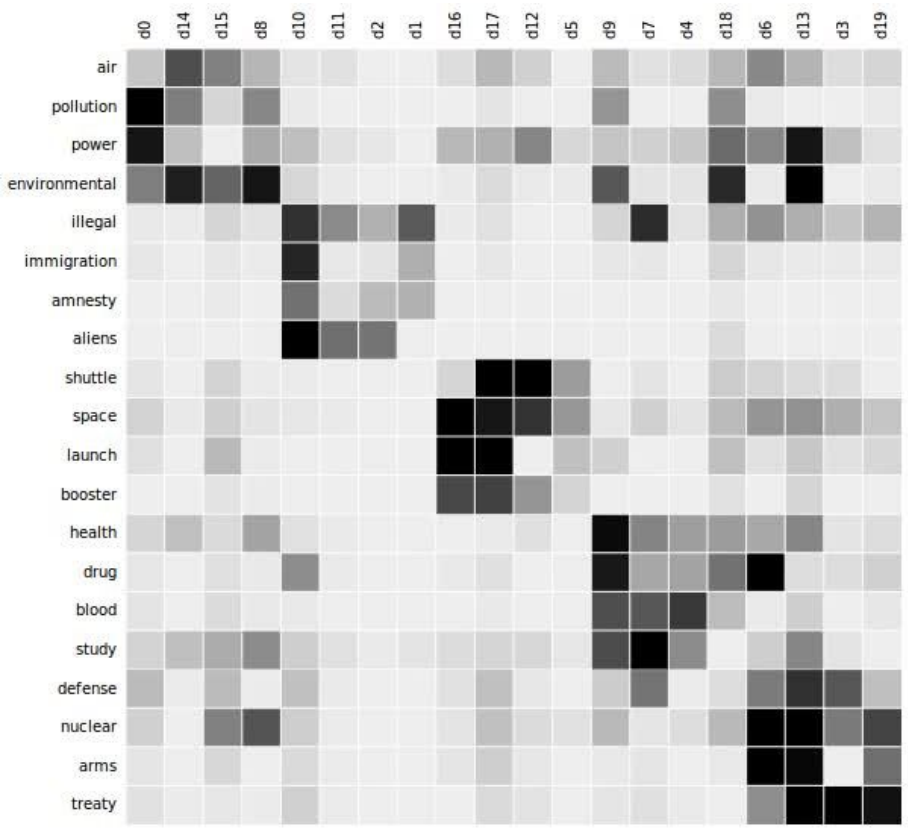
“Universidade de São Paulo”

ou

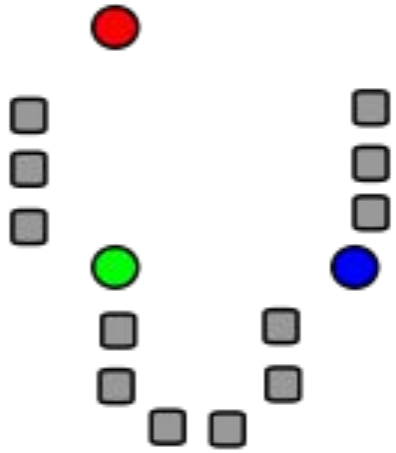
“Univ. de SP”

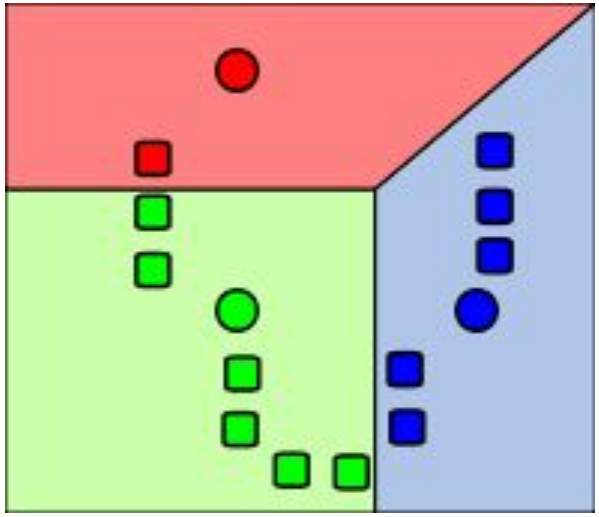
Análise Semântica Latente (*LSA*)



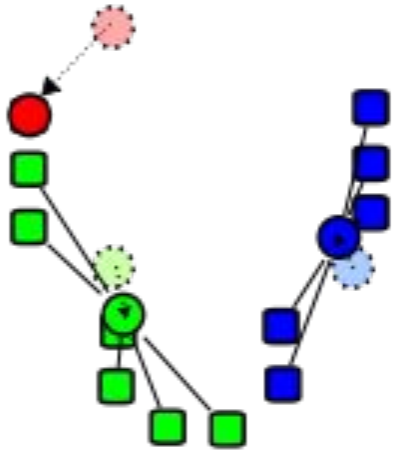


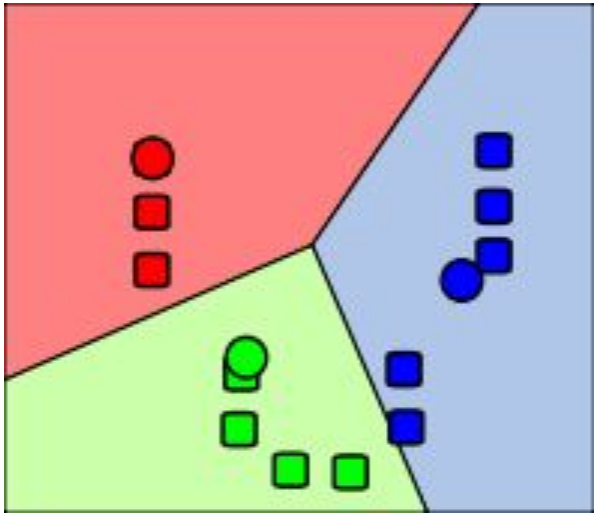
k-Means





k-Means

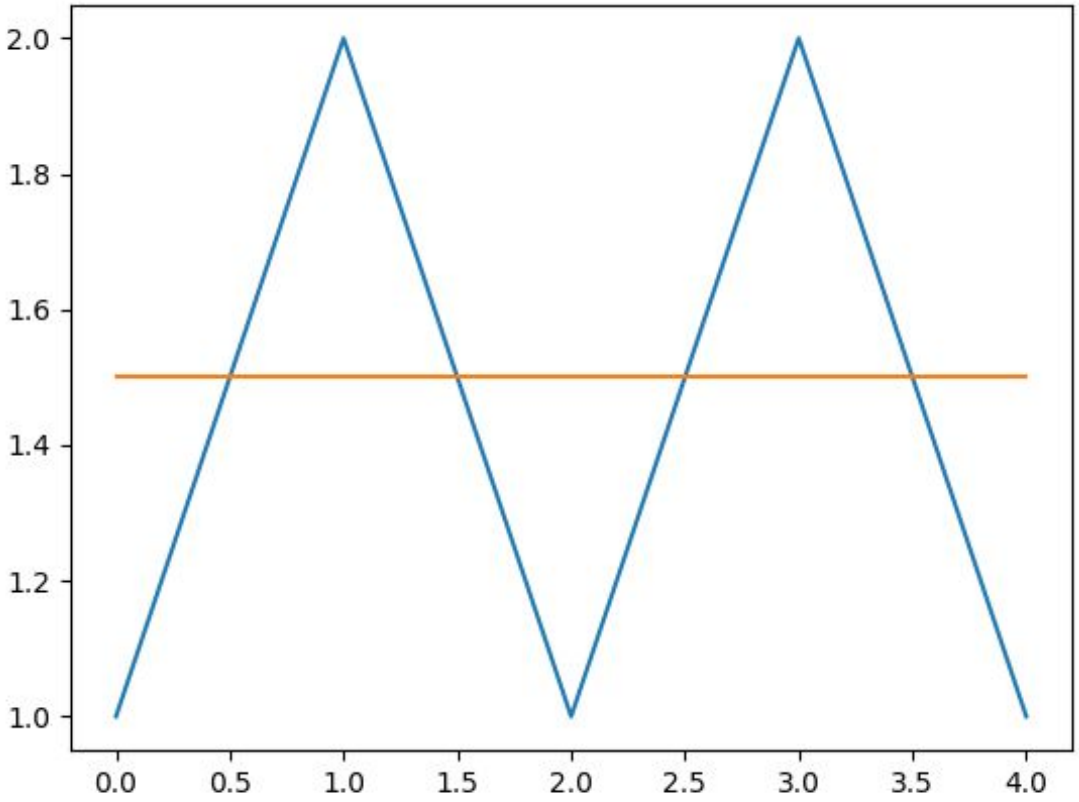




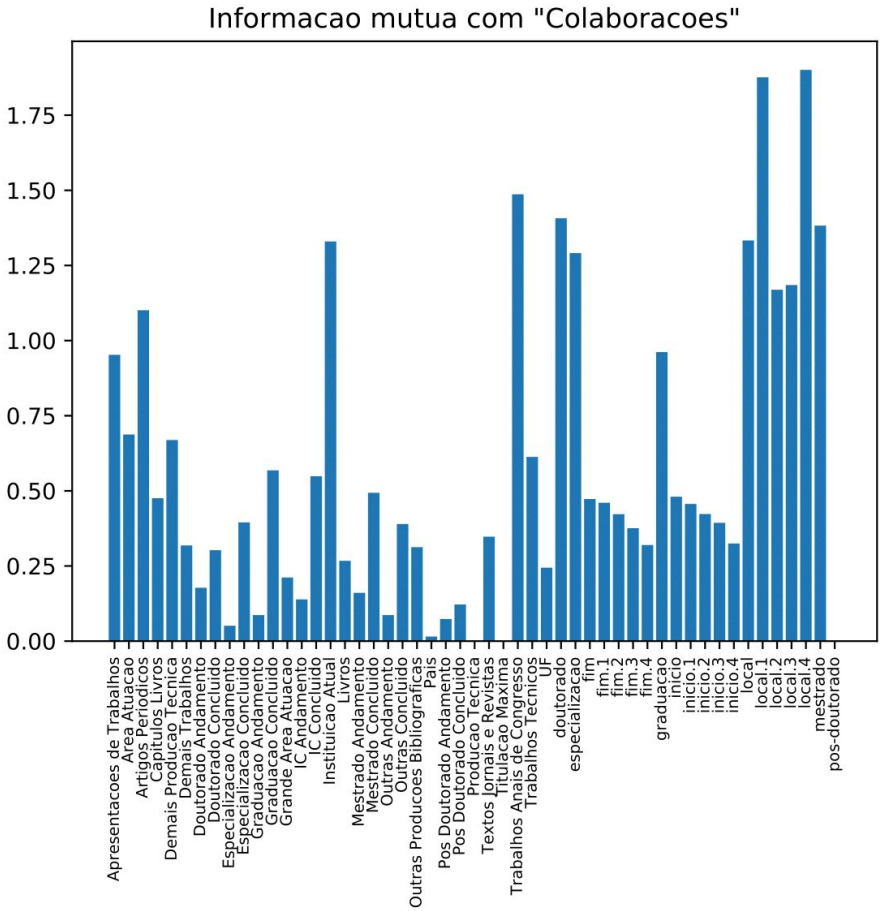
Informação Mútua

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Informação Mútua



Informação Mútua



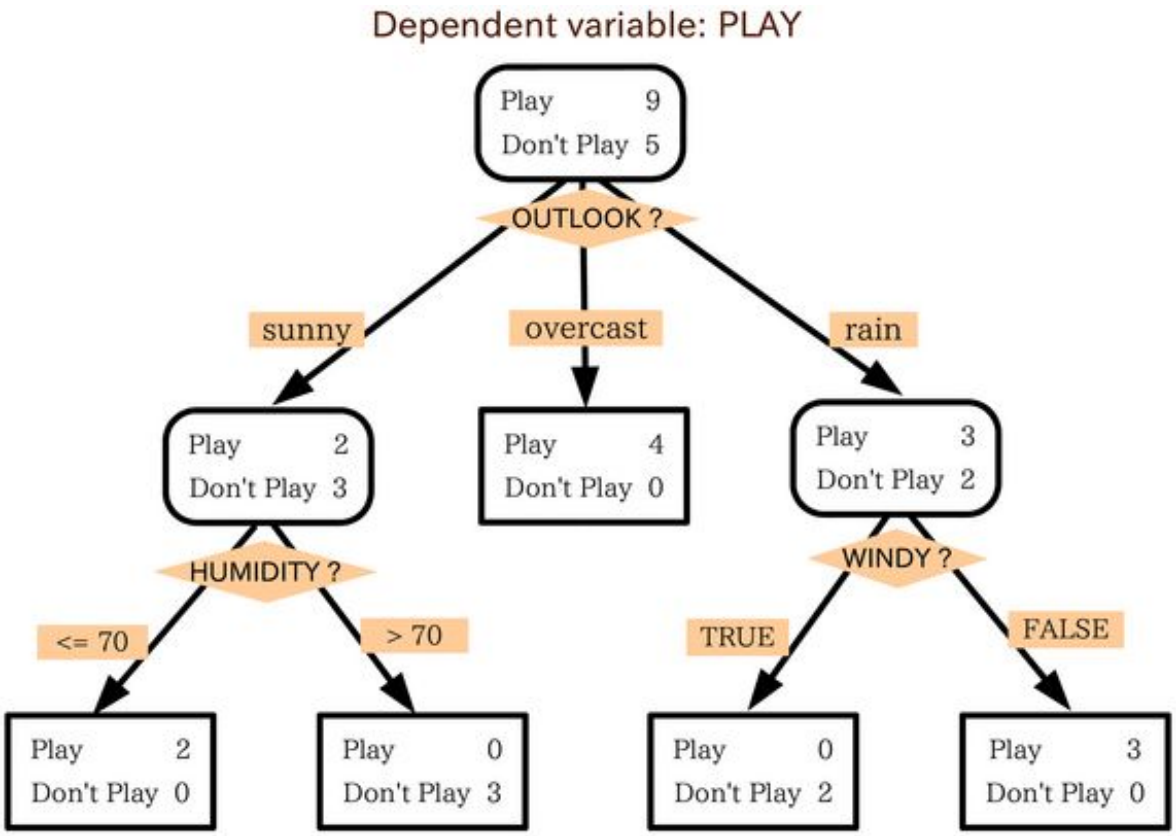
Resultado

8.180 currículos

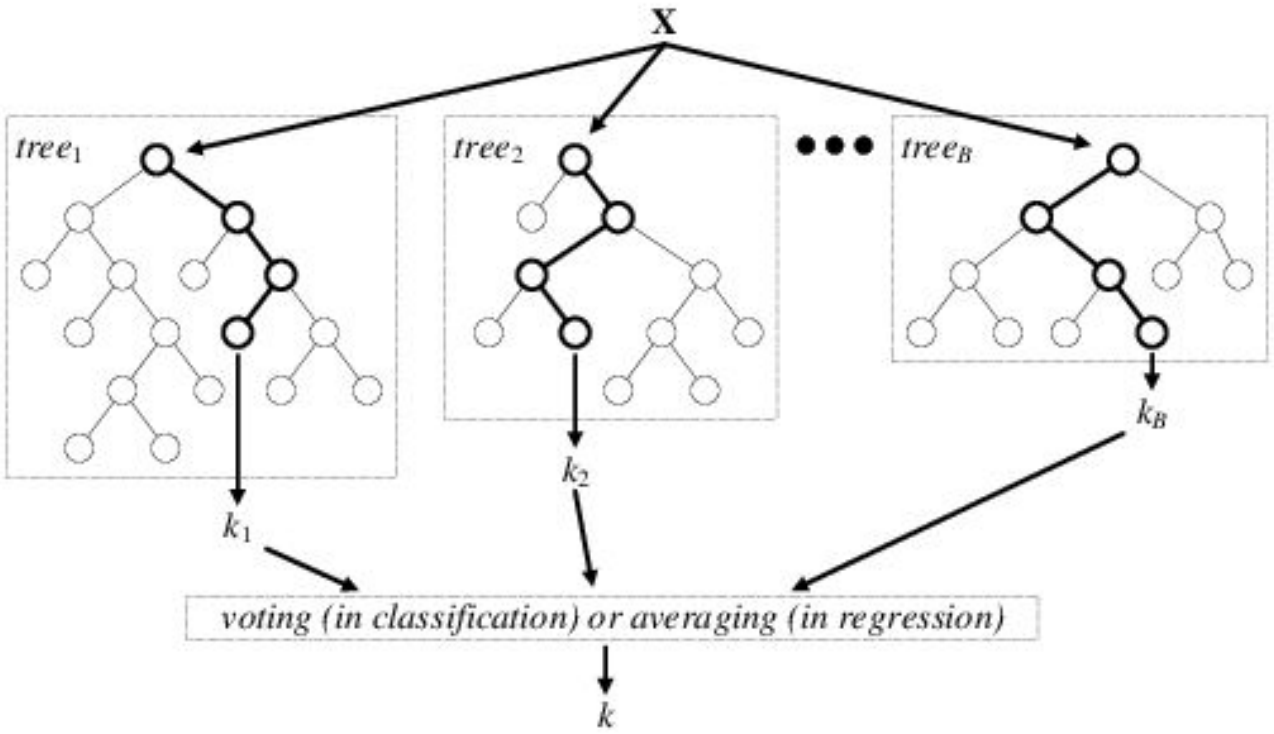
58.689 colaborações

Modelo de Mineração de Dados

Árvores de Decisão



Florestas Aleatórias

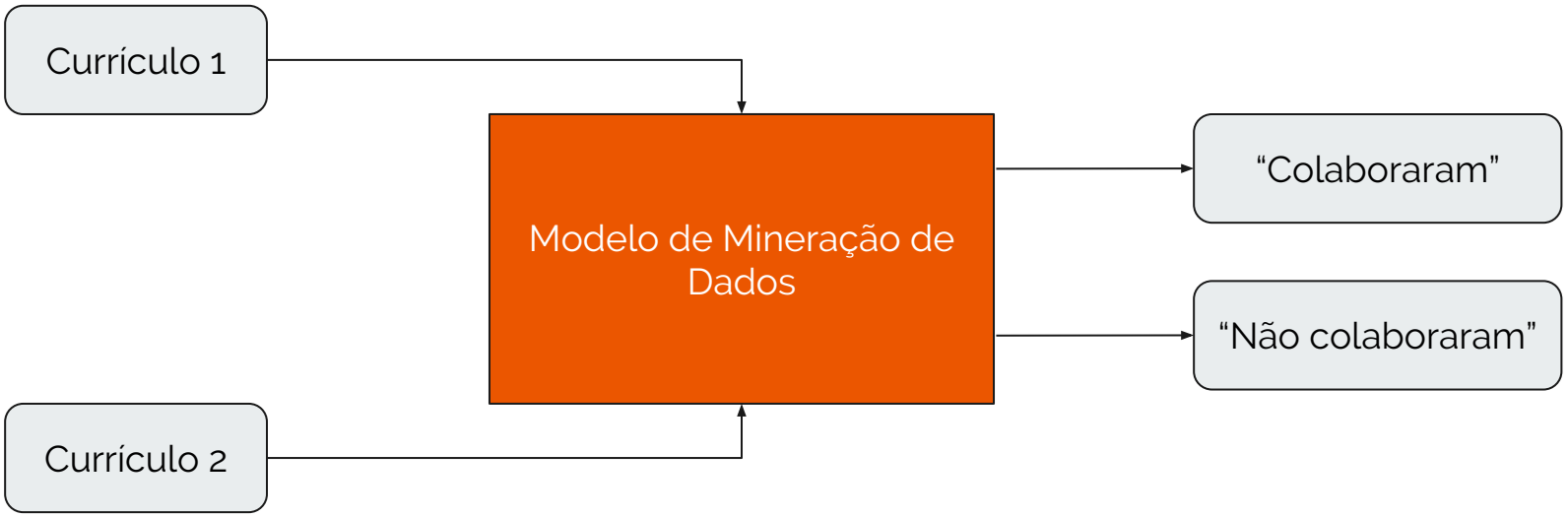


Experimentos

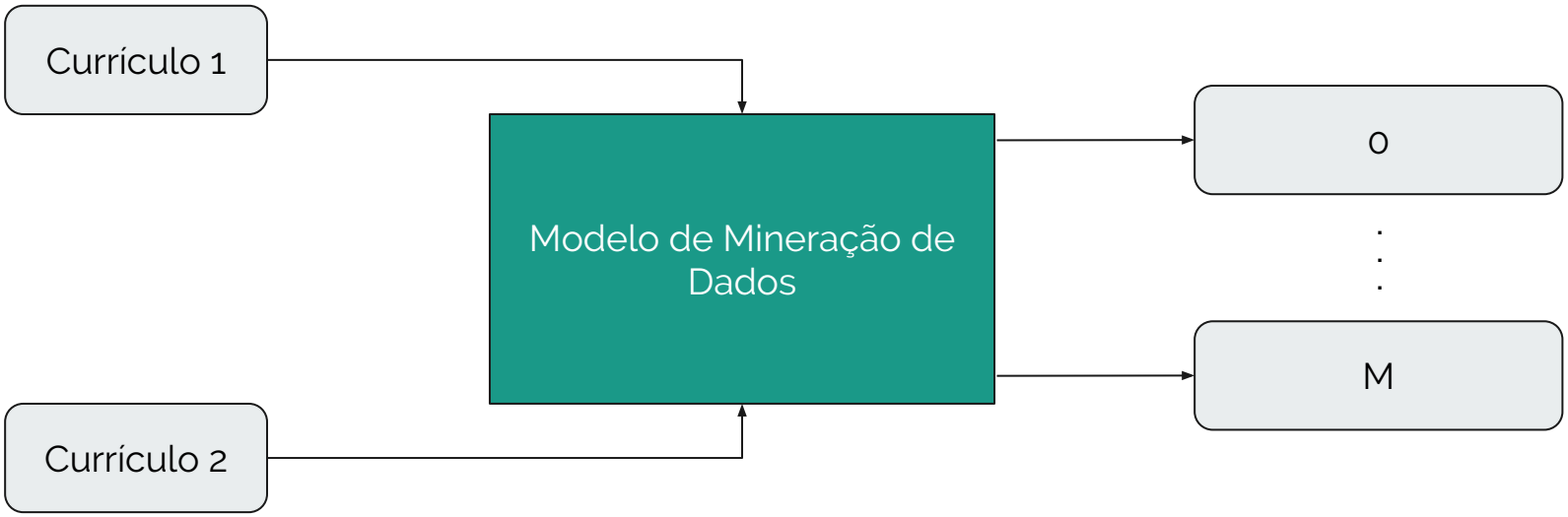
Divisões da base

- Estratificada em colaborações;
- Treino - 60%;
- Teste - 40%.

Modalidade *binary*



Modalidade *standard*



Resultados

Acurácia

Binary

$$86,70\% \pm 0,18$$

Standard

$$78,34\% \pm 0,16$$

Conclusão

- O modelo proposto é preciso o suficiente para ser utilizado na prática;
- Limitações:
  - Tratamento de dados faltantes;
  - Não generaliza para termos não presentes no conjunto de dados;
- Melhorias possíveis:
  - Mineração de texto mais robusta;
  - Integrar o processo de colheita de uma nova versão da base às demandas do modelo.

Dúvidas?