Universidade de Brasília
Departamento de Estatística

~~Proposta de~~ agrupador de processos de controle concentrado

Bruno Gondim Toledo

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Bruno Gondim Toledo

Proposta de agrupador de processos de controle concentrado

Orientador(a): Profª Drª Thais Carvalho Valadares Rodrigues.

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Sumário

1 Introdução

Modelagem estatística ~~vem se tornando~~ essencial em diversos setores, inclusive no poder Judiciário, visto que tem sido do interesse dos próprios tribunais saber indicadores e estatísticas relacionadas ao próprio tribunal, além de agilizar processos e direcionar os recursos humanos para tarefas menos mecânicas. ~~E~~ a modelagem surge neste contexto para agregar valor às decisões estratégicas deste poder. Com o advento de novas técnicas computacionais, é possível agrupar e classificar textos do contexto jurídico (FREITAS, 2024).

O propósito deste trabalho é formular um agregador de processos de controle concentrado do Supremo Tribunal Federal (STF). Compõem o controle concentrado ~~as~~ classes processuais ~~as~~ Ações Diretas de Inconstitucionalidade (ADIs), as Ações Declaratórias de Constitucionalidade (ADCs), as Ações Diretas de Inconstitucionalidade por Omissão (ADOs) e as Arguições de Descumprimento de Preceito Fundamental (ADPFs).

Para a construção deste agregador, pretende-se vetorizar as petições dos processos de controle concentrado com a técnica de ~~*Embedding*~~ utilizada em *Text clustering applied to data augmentation in legal context~~s~~* (FREITAS, 2024), com posterior aplicação de métodos de agrupamento destes vetores. Uma etapa anterior ~~ao *Embedding*~~ é o pré-processamento do texto obtido, com~~o~~ remoção de *stopwords*, pontuações, termos legais, números, espaços em branco FREITAS (2023).

Por conta da natureza cíclica dos processos que compõem o acervo do ~~tribunal~~, não é esperado a formulação de classes rígidas ~~que poderiam compor~~ este agrupamento, sendo somente ~~possível~~ encontrar similaridades e dissimilaridades entre os processos em atual tramitação à fim de agrupá-los de forma a facilitar o fluxo de trabalho do tribunal. Portanto, a proposta deste trabalho é estudar técnicas de agrupar estes processos de forma funcional e contínua, fornecendo assim um produto para o tribunal, ~~possivelmente na forma de um *Shiny*~~, a fim de ser utilizado pelos analistas que hoje buscam manualmente esses agrupamentos.

2 Objetivos

2.1 Objetivo Geral

Elaborar ~~aplicação que agrupe~~ processos de controle concentrado de ~~forma a fornecer~~ subsídios aos responsáveis pelo ~~seu~~ encaminhamento, visando reduzir trabalho mecânico ~~de~~ analisar e encaminhar processos semelhantes aos setores responsáveis ~~do tribunal~~.

2.2 Objetivos Específicos

- ~~objetivo 1 - Entender o objeto de estudo e a demanda por uma aplicação~~;
- ~~objetivo 2 - Entrevistar os analistas do tribunal a fim de alinhar a demanda com a possibilidade teórica e prática, bem como estabelecer um fluxo em conformidade com demais aplicações existentes no tribunal;~~
- ~~objetivo 3 - Estabelecer um recorte temporal que seja compatível com a demanda e os dados a serem analisados;~~
- ~~objetivo 4 - Coletar os dados definidos no recorte temporal. Conforme definido na Introdução, os dados serão as petições iniciais dos processos;~~
- ~~objetivo 5 -~~ Processar os dados, transformando-~~os de textos em~~ PDF para vetores numéricos;
- ~~objetivo 6 - Analisar, estudar e transformar os dados para agrupamento~~;
- ~~objetivo 7 -~~ Aplicar tecnologias de agrupamento ~~aos~~ dados;
- ~~objetivo 8 - Comparar~~ os agrupamentos fornecidos pela aplicação com as reais demandas do tribunal.

O ~~objetivo 5~~ será realizado utilizando o fluxo do tribunal, ou seja, os modelos já existentes neste para ocerização, processamento de texto e ~~*embedding*~~, a fim de tornar a aplicação comparável com as demais existentes no tribunal. ~~Os objetivos 6, 7 e 8 serão realizados de forma recursiva, ou seja, refinar o método de agrupamento sucessivamente de forma que a tecnologia de fato auxilie no agrupamento de processos, sendo ao menos igual ou idealmente superior ao trabalho manual atualmente realizado. Um dos possíveis ganhos desta aplicação é a priorização de pauta de determinados processos.~~

3 Metodologia

A metodologia aplicada para processamento e conversão de dados textuais em vetores numéricos para posterior classificação será a mesma aplicada por FREITAS (2024), consistindo nas etapas de ~~*Text cleaning*, *Embedding* e *Data organization*~~.

3.1 Vetorização

Vetorização, ou *Embedding*, é uma técnica geralmente aplicada na área de processamento de linguagem natural. Em poucas palavras, é um procedimento de transformar texto em um vetor numérico.

A ideia por trás da vetorização é transformar os dados originais, texto, em um vetor numérico que mapeie e preserve a carga semântica do texto original. Sem embargos, é possível rapidamente identificar a utilidade desta técnica — visto que, para utilização de técnicas e algoritmos estatísticos, em geral buscamos dados numéricos, e não textuais — quando intencionamos aplicar alguma técnica para modelagem destes dados. Diversas técnicas para este procedimento vem sendo estudadas, implementadas e utilizadas, como por exemplo o método *Term Frequency — Inverse Document Frequency* (TF-IDF) (Salton e Buckley, 1988, apud FREITAS (2024)), *Word2vec* (Mikolov et al, 2013, apud FREITAS (2024)) e *doc2vec* (Le e Mikolov, 2014, apud FREITAS (2024)). Este último busca a obtenção do vetor representando o texto original via o ~~treinamento do vetor do texto por meio de atualização dos pesos em conjunto do vetor de palavras que formam coletivamente o texto~~ (FREITAS, 2024). Mais detalhes do funcionamento destes métodos podem ser encontrados em FREITAS (2024).

Para esta aplicação, utilizaremos o método *doc2vec*, em conformidade com a rotina de vetorização de textos do Supremo Tribunal Federal.

3.2 Medidas de similaridade e dissimilaridade

Para agrupar os processos, será necessário estudar medidas de similaridade e dissimilaridade. Artes e Barroso (2023) argumenta que para medidas de similaridade, quanto maior o valor, maior será a semelhança entre objetos, enquanto que para medidas de dissimilaridade, quanto maior o valor, mais diferente serão os objetos. Dado que estarei trabalhando com vetores numéricos fornecidos pelo processamento dos dados ~~realizado anteriormente~~, um dos caminhos possíveis a seguir é de pensar em distâncias estatísticas.

3.2.1 Distância estatística

Uma medida de distância estatística pode ser definida simplesmente como a distância entre um ponto arbitrário P de um segundo ponto arbitrário Q, em qualquer espaço n-dimensional.

Conforme definido por JOHNSON e WICHERN (2007), qualquer medida de distância $d(P, Q)$ entre os pontos P e Q é válida se atender as seguintes propriedades, sendo R um terceiro ponto intermediário qualquer:

- $d(P, Q) = d(Q, P)$
- $d(P, Q) > 0 \; se \; P \neq Q$
- $d(P, Q) = 0 \; se \; P = Q$
- $d(P, Q) \leq d(P, R) + d(R, Q)$ (Desigualdade triangular)

A última propriedade será evocada quando definirmos a Distância do cosseno.

3.2.2 Distância Euclidiana

A distância Euclidiana é uma consequência direta do Teorema de Pitágoras. Segundo definido por JOHNSON e WICHERN (2007), considerando os pontos arbitrários P com coordenadas $P = (x_1, x_2, ..., x_p)$ e Q com coordenadas $Q = (y_1, y_2, ..., y_p)$ de uma hiperesfera p-dimensional, a distância reta entre eles é dada por:

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \tag{3.2.1}$$

O que define a distância Euclidiana.

Note que esta definição para P, Q pode ser expandida para a situação em que estes são vetores X_{ik}, X_{kj}, i, j amostras com j dimensões. Utilizando a notação de Artes e Barroso (2023):

$$d_{ik} = \sqrt{(x_i - x_k)^T (x_i - x_k)} = \sqrt{\sum_{j=1}^{p} (X_{ik} - X_{kj})^2} \tag{3.2.2}$$

Note que 3.2.1 é um caso particular desta quando $j = 1$.

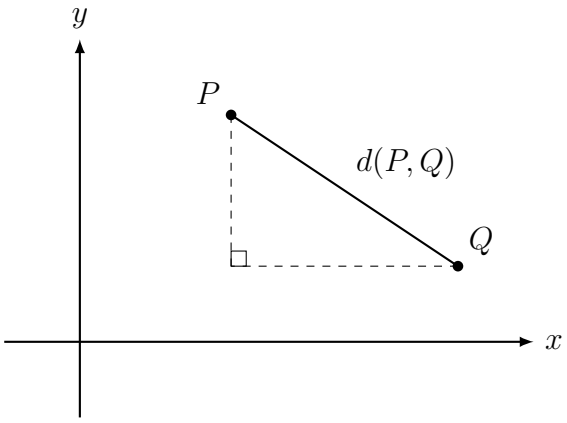

Figura 1: Distância Euclidiana

3.2.3 Distância de quarteirão

Seja $x_i = (X_{i1}, ..., X_{ip})^T$ vetor de observações ~~de i amostras~~, $i = 1, ..., n$, no qual X_{ij} ~~o~~ valor assumido pela variável X_j ~~para a~~ amostra i. Artes e Barroso (2023) define a distância de quarteirão (ou *Manhattan*) entre a i-ésima amostra e a ~~j~~-ésima amostra como

$$d_{ik} = \sum_{j=1}^{p} |X_{ij} - X_{kj}|$$

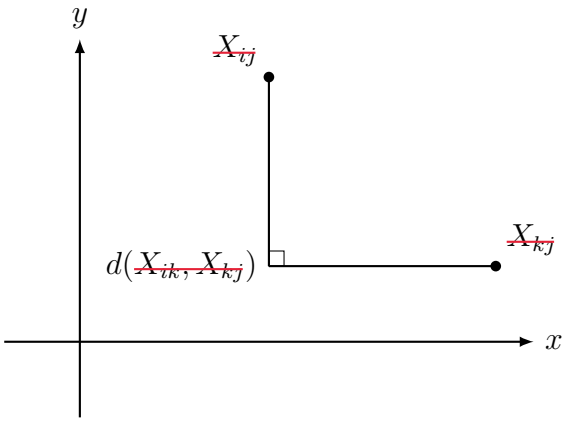

Figura 2: Distância de quarteirão

3.2.4 Distância de Mahalanobis

(ARTES; BARROSO, 2023) Seja $\mathbf{x} = (X_1, ..., X_p)^T$ vetor aleatório com $\mathbb{E}(\mathbf{x}) = \mu = (\mu_1, ..., \mu_p)^T$, e $Cov(\mathbf{x}) = \mathbf{\Sigma}$.

A distância de Mahalanobis ao quadrado é dada por

$$D_M^2(\mathbf{x}, \mu) = (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \tag{3.2.3}$$

Seja $i = i, ..., n$ unidades amostrais de dimensão p. Logo, o vetor $\mathbf{x} = (\mathbf{x}_i, ..., \mathbf{x}_p)$ define a amostra. Seja $\bar{\mathbf{x}}$ vetor de média desta amostra, e $\mathbf{S}$ matriz de covariância amostral. Logo, a distância de um ponto i em relação centróide da amostra será

$$D_M^2(\mathbf{x_i}, \bar{\mathbf{x}}) = (\mathbf{x_i} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}}) \tag{3.2.4}$$

Pela definição, necessitaríamos saber os valores μ e $\mathbf{\Sigma}$. Artes e Barroso (2023) argumenta que em termos práticos, isto não ~~ocorre~~. Portanto, sugere utilizar os estimadores usuais $\bar{\mathbf{x}}$ e $\mathbf{S}$, respectivamente, dando suporte para a definição 3.2.4. Os resultados passam a ser aproximados, com melhor desempenho para grandes amostras.

3.2.5 Distância de cossenos

Seja $\mathbf{A}$ e $\mathbf{B}$ vetores não nulos. Então, pela Lei dos cossenos:

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}| \cos\theta \tag{3.2.5}$$

Ricardo e Berthier (2011) ~~define~~ que podemos utilizar para avaliar o grau de similaridade entre os vetores $\mathbf{A}$ e $\mathbf{B}$ a correlação entre eles. E para quantificar esta correlação, podemos utilizar o cosseno do ângulo entre estes vetores. ~~De~~ 3.2.5, temos que esta similaridade será:

$$\begin{aligned} \text{similaridade}(\mathbf{A}, \mathbf{B}) = \cos\theta &= \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \\ &= \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{A}_i^2} \cdot \sqrt{\sum_{i=1}^{n} \mathbf{B}_i^2}} \end{aligned}$$

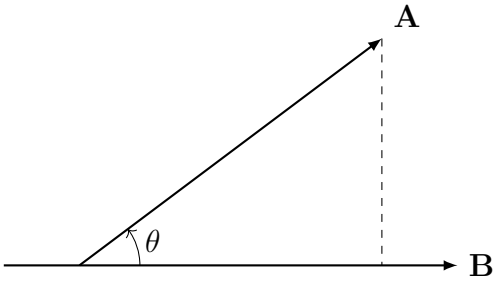

Figura 3: O cosseno como medida de similaridade entre $\mathbf{A}$ e $\mathbf{B}$

Se considerarmos **A** e **B** os vetores numéricos das petições de dois processo ~~vetorizadas~~, conseguimos assim calcular uma medida de similaridade entre eles.

Note que a distância do cosseno não respeita a desigualdade triangular definida em 3.2.1, portanto não pode ser considerada uma medida de distância estatística naquelas definições.

3.3 Métodos de agrupamento hierárquicos

Métodos hierárquicos buscam dividir ou agrupar um conjunto de dados em uma sequência de partições aninhadas (BORRIES; WANG, 2009), segundo alguma medida de similaridade (ou dissimilaridade). Estes dividem-se em dois grupos.

3.3.1 Aglomerativo

Este método se inicia com uma observação em cada grupo, e a cada iteração as observações são agrupadas ao seu par mais similar de acordo com alguma medida de similaridade ou dissimilaridade, até que todas as observações estejam em um único grupo. Alguns exemplos de algoritmos que executam este tipo de agrupamento são:

1. Ligação Simples: Une os elementos mais próximos;
2. Ligação Completa: Une os elementos mais distantes;
3. Ligação média: Une os elementos de acordo com suas respectivas medóides;
4. Ward: Faz novos grupos baseado na ~~menos~~ soma de quadrados ~~do erro~~ de todas as possíveis combinações de pares.

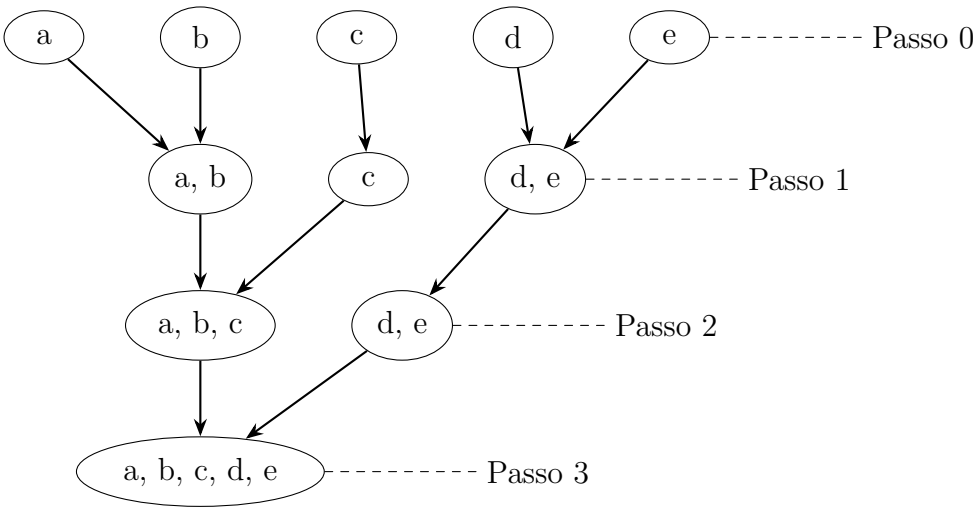

Figura 4: Fluxograma do algoritmo aglomerativo

Para mais informações, consultar Kaufman e Rousseeuw (1990)

3.3.2 Divisivo

Faz o caminho inverso do aglomerativo, ou seja, inicia o procedimento com todas as observações em um único grupo, e a cada iteração as observações são separadas utilizando alguma medida de similaridade ou dissimilaridade, até que cada observação forme um grupo contendo apenas ela. ~~Primeiro~~ apresentado por Kaufman e Rousseeuw (1990) com o algoritmo DIANA (*Divisive Analysis* - Análise divisiva)

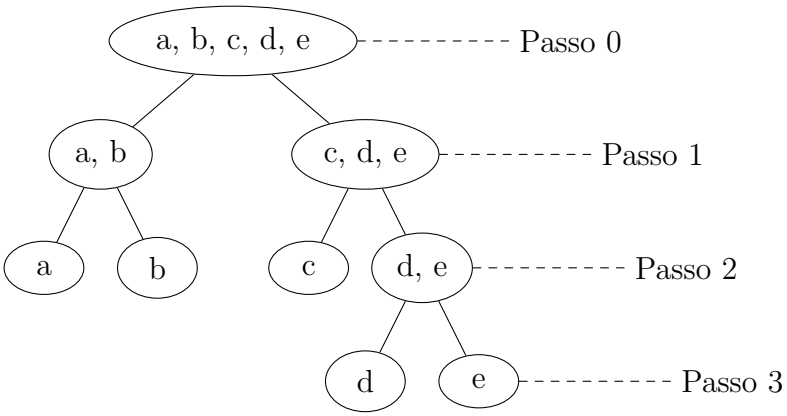

Figura 5: Fluxograma do algoritmo divisivo

3.4 Métodos de agrupamento não hierárquicos

Os métodos não hierárquicos buscam encontrar uma partição dos n elementos em g grupos. Estes grupos são representados por suas medóides e possuem forma convexa (elipsoide).

3.4.1 k-médias

k-médias, geralmente evocado como *k-means*, ou ainda k-medóides (ARTES; BARROSO, 2023) é um algoritmo que busca agregar pontos em K grupos, de forma a minimizar a soma dos quadrados das distâncias euclidianas entre o centro (meio) de cada grupo~~,~~ e ~~d~~os pontos que o compõem~~.~~ (MORETTIN; SINGER, 2021)

Seja k o número de grupos em que se deseja agrupar os dados. Seja $\mathbf{x} = \mathbf{x}_1, ..., \mathbf{x}_n$ vetor de amostra.

O procedimento realizado pelo algoritmo pode ser descrito como:

- Fixar um número de k grupos (O usuário precisa especificar este número);

- ~~Então, o algoritmo~~ aloca os elementos aleatoriamente aos k grupos, e calcula a medóide $\bar{\mathbf{x}}_k$ de cada grupo;
- ~~O algoritmo calcula~~ a distância Euclidiana de cada ~~item~~ ao medóide do grupo em que está alocado:

$$d_{i,\bar{x}_k} = \sum_{k=1}^{k} \sum_{g_{(i)}=k} (\mathbf{x}_i - \bar{x}_k)^T (\mathbf{x}_i - \bar{x}_k) \tag{3.4.1}$$

  em que $\bar{x}_k$ é a medóide do grupo k, e $g_{(i)}$ é o grupo que contém $\mathbf{x}_i$

- ~~Realoca~~ cada item $\mathbf{x}_i$ ao grupo com medóide mais próximo tal que reduza $d_{i,\bar{x}_k}$. Então, calcula os novos valores de $\bar{x}_k$.
- Repetir os dois últimos passos de forma iterativa, até estabilizar o valor de $d_{i,\bar{x}_k}$.

Este algoritmo é especialmente útil quando temos alguma ideia de quantos k grupos desejamos separar os dados.

3.5 Conjunto de dados

O conjunto de dados ~~selecionado é o que compõe o escopo do trabalho, ou seja~~, a petição inicial dos processos de controle concentrados tramitados ou em tramitação entre 1º de Janeiro de 2016 e 10 de Abril de 2024 ~~- Quando o recorte temporal fora definido~~. Os dados são públicos e encontram-se disponíveis em Corte aberta.

Optou-se pelo recorte à partir de 2016 a fim de parear o recorte deste estudo com o recorte estabelecido por outras aplicações do Supremo Tribunal Federal, como a RAFA2030.

A aplicação da ocerização e pré-processamento do texto, bem como algoritmo de vetorização foram fornecidas pela Secretaria de Tecnologia da Informação (STI) do tribunal para a manufatura deste trabalho.

Como os agrupamentos são construídos a partir dos dados que compõem o acervo atual do tribunal em uma determinada data, será testada a técnica em diversos recortes temporais dos dados obtidos, considerando em cada teste apenas os dados que ~~compunham~~ o acervo na data fixada de cada teste.

3.6 Visualização de dados

3.6.1 Dendograma

(EVERITT; SKRONDAL, 2010) Dendograma é um diagrama geralmente utilizado para ilustrar a série de etapas executada por um método hierárquico, representando os passos de agrupamento do algoritmo. A altura do eixo y representa alguma medida de distância entre os agrupamentos.

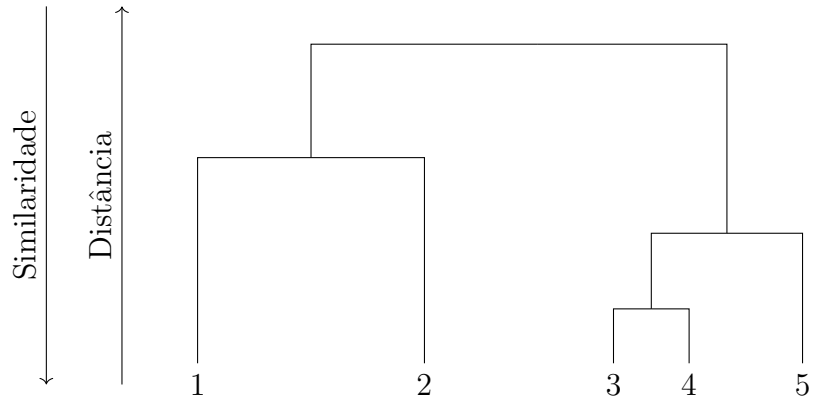

Figura 6: Exemplo simples de um dendograma

3.6.2 t-SNE

Agrupamento estocástico de vizinhança com distribuição t-student — *t-distributed stochastic neighbor embedding (t-SNE)*, tradução livre — é uma técnica não linear de visualização de dados multidimensionais em duas (ou três) dimensões. Esta técnica busca apresentar a estrutura de todo um conjunto de dados multidimensional com escalas diversas em um só gráfico.

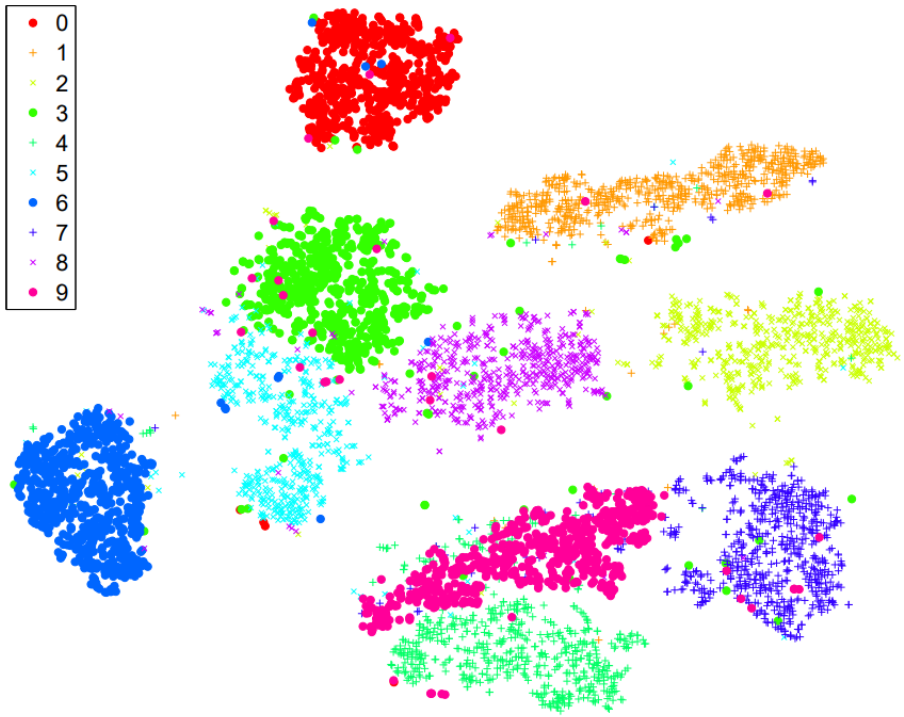

Figura 7: O algoritmo t-SNE aplicado aos dados MNIST
Fonte: Maaten e Hinton (2008)

O conjunto de dados MNIST (LECUN et al., 1998) se tornou "padrão ouro" nas diversas áreas de reconhecimento de padrões, como análise multivariada clássica, processamento de linguagem natural, etc.

A visualização *t-SNE* não busca modelar o conjunto de dados, mas sim apresentar de forma visual o resultado produzido pelo algoritmo utilizado. Por vezes, técnicas clássicas de visualização multivariada produzem visualizações de difícil interpretação, com interpolações dos agrupamentos formados na projeção dos resultados em duas dimensões. Portanto, o *t-SNE* busca separar no gráfico estes agrupamentos formados, esclarecendo a estrutura dos resultados do procedimento de forma visualmente compreensível.

Para mais detalhes do funcionamento da técnica, consultar Maaten e Hinton (2008).

4 Cronograma

As atividades a serem desenvolvidas durante o Trabalho de Conclusão de Curso em 1/2021 são:

1. Escolha do tema a ser abordado.
2. Levantamento de bibliografia relacionada ao tema.
3. Definição do recorte temporal com a AAJ do STF.
4. Solicitação dos dados para a STI do STF.
5. Solicitação dos algoritmos à STI do STF.
6. Revisão de literatura.
7. Desenvolvimento da proposta de projeto.
8. Análise preliminar do banco de dados.
9. Entrega da proposta de projeto.
10. Elaboração da apresentação da proposta.
11. Manipulação do banco de dados.
12. Análise do banco de dados.
13. Elaboração do relatório parcial.
14. Entrega do relatório parcial a Professora Orientadora.
15. Correção do relatório parcial.
16. Entrega do relatório parcial a banca.
17. ~~Elaboração do relatório parcial.~~
18. ~~Entrega do relatório parcial ao Professor Orientador(a).~~
19. ~~Correção do relatório parcial.~~
20. ~~Entrega do relatório parcial para a banca.~~
21. Desenvolvimento do modelo e da aplicação.
22. Elaboração do relatório final.

23. Entrega do relatório final ao Professor Orientador(a).

24. Correção do do relatório final.

25. Entrega do relatório final para a banca.

Tabela 1: Cronograma

	1/2024					2/2024				
	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubroro	Novembro	Dezembro
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										

Referências

ARTES, R.; BARROSO, L. P. *Métodos multivariados de análise estatística*. [S.l.]: São Paulo: Blucher, 2023.

BORRIES, G. von; WANG, H. Partition clustering of high dimensional low sample size data based on p-values. *Computational statistics data analysis, v. 53, n. 12, p. 3987-3998*, 2009.

EVERITT, B.; SKRONDAL, A. *The cambridge dictionary of statistics.* [S.l.]: Cambridge University Press, 2010. v. 4.

FREITAS, L. J. G. e. a. Catboost algorithm application in legal texts and un 2030 agenda. *Revista de Informatica Teórica e Aplicada - RITA - ISSN 2175-2745. Vol. 30, Num. 02 (2023) 51-58*, 2023.

FREITAS, L. J. G. e. a. Text clustering applied to data augmentation in legal contexts. *arXiv preprint arXiv:2404.08683*, 2024.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis.* [S.l.]: 6. ed.[S.l.]:Prentice Hall, 2007.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 1990.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. 11, 2008.

MORETTIN, P. A.; SINGER, J. M. *Estatística e Ciência de Dados*. [S.l.]: LTC, 2021.

RICARDO, B.-Y.; BERTHIER, R.-N. *Modern information retrieval: the concepts and technology behind search*. [S.l.]: New Jersey, USA: Addi-son-Wesley Professional, 2011.