



**Universidade de Brasília
Departamento de Estatística**

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

Agrupador de processos de controle concentrado

Orientador(a): Prof^a Dr^a 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	4
2 Objetivos	5
2.1 Objetivo Geral	5
2.2 Objetivos Específicos	5
3 Metodologia	6
3.1 Vetorização	6
3.2 Medidas de similaridade e dissimilaridade	6
3.2.1 Distância Euclidiana	7
3.2.2 Distância de quarteirão	8
3.2.3 Distância de Mahalanobis	9
3.2.4 Distância de cossenos	9
3.3 Métodos de agrupamento hierárquicos	10
3.3.1 Aglomerativo	10
3.3.2 Divisivo	11
3.4 Métodos de agrupamento não hierárquicos	12
3.4.1 k-médias	12
3.5 Conjunto de dados	12
3.6 Visualização de dados	13
3.6.1 Dendograma	13
3.6.2 t-SNE	14
4 Cronograma	16

1 Introdução

Modelagem estatística tornou-se 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. 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, conforme descrito por Freitas et al. (2024).

O propósito deste trabalho é formular um agregador de processos de controle concentrado do Supremo Tribunal Federal (STF). Compõem o controle concentrado as seguintes classes processuais: 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 vetorização (do inglês, *embedding*) utilizada em "*Text clustering applied to data augmentation in legal contexts*" (Freitas et al, 2024), com posterior aplicação de métodos de agrupamento destes vetores. Uma etapa anterior à vetorização é o pré-processamento do texto obtido, como a remoção de *stopwords*, pontuações, termos legais, números, espaços em branco, conforme discutido em Freitas et al. (2023).

Por conta da natureza cíclica dos processos que compõem o acervo do STF, não é esperado a formulação de classes rígidas para este agrupamento, sendo somente necessário 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. Por exemplo, ao agrupar processos semelhantes, estes podem ser analisados conjuntamente, facilitando o fluxo de trabalho do STF. Além disso, o agrupador pode facilitar uma priorização de pauta de determinados processos. Portanto, a proposta deste trabalho é estudar técnicas de agrupar estes processos de forma funcional e contínua, fornecendo assim um produto para o tribunal, a fim de ser utilizado pelos analistas que hoje buscam manualmente esses agrupamentos.

2 Objetivos

2.1 Objetivo Geral

Elaborar um agrupador de processos de controle concentrado para os processos em tramitação no Supremo Tribunal Federal. O agrupador fornecerá subsídios aos responsáveis pelo encaminhamento dos processos que chegam ao STF, visando reduzir o trabalho mecânico dos servidores do tribunal que atualmente necessitam analisar e encaminhar processos semelhantes aos setores responsáveis.

2.2 Objetivos Específicos

- Processar os dados utilizando técnicas de Processamento de Linguagem Natural (PLN), transformando as petições iniciais dos processos em PDF (dados de entrada) para vetores numéricos;
- Comparar técnicas;
- Aplicar e comparar tecnologias de agrupamento dos dados;
- Avaliar os agrupamentos fornecidos pela aplicação com as reais demandas do tribunal.
- Estudar técnicas de processamento de linguagem natural, análise multivariada e visualização de dados.

O processamento inicial dos dados (transformação dos textos em vetores numéricos) será realizado utilizando o fluxo do tribunal, ou seja, os modelos já existentes neste para ocerização, processamento de texto e vetorização, a fim de tornar a aplicação comparável com as demais existentes no tribunal.

3 Metodologia

A metodologia aplicada para processamento e conversão de dados textuais em vetores numéricos para posterior classificação será a mesma aplicada por Freitas et al. (2024), consistindo nas etapas de limpeza de dados, vetorização e organização dos dados. Após o processamento inicial, os dados serão agrupados segundo alguma medida de similaridade. Métodos de agrupamento hierárquico e não hierárquico serão explorados.

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 et al, 2024), *Word2vec* (Mikolov et al, 2013, apud Freitas et al, 2024) e *doc2vec* (Le e Mikolov, 2014, apud Freitas et al, 2024). Este último busca a obtenção do vetor numérico representando o texto original via treinamento de uma rede neural. Nessa rede, os parâmetros do modelo (ou pesos) são atualizados buscando prever a próxima palavra dado o seu contexto (palavras ao redor) e dado um vetor de palavras que formam coletivamente o texto, conforme descrito em Freitas et al. (2024). Mais detalhes do funcionamento destes métodos podem ser encontrados em Freitas et al. (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) argumentam 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 inicial dos dados (3.1),

um dos caminhos possíveis a seguir é de pensar em distâncias estatísticas.

Uma medida de distância estatística pode ser definida simplesmente como a distância entre um ponto arbitrário P e 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 (seção 3.2.4). A seguir serão apresentadas algumas medidas de distância que poderão ser exploradas, sendo elas as distâncias euclidiana, de quarteirão, de Mahalanobis e de cossenos.

3.2.1 Distância Euclidiana

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

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

A Figura 1 ilustra geometricamente a distância entre dois pontos \mathbf{P} e \mathbf{Q} no \mathbb{R}^2 .

Note que esta definição para P, Q pode ser expandida para a situação em que estes são vetores X_i, X_k , i, k amostras com p dimensões. Utilizando a notação de Artes e Barroso (2023):

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

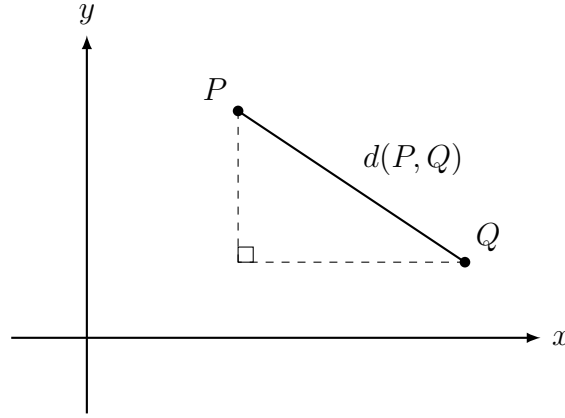


Figura 1: Distância Euclidiana entre os pontos **P** e **Q**

3.2.2 Distância de quarteirão

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

$$d_{ik} = \sum_{j=1}^p |X_{ij} - X_{kj}| \quad (3.2.3)$$

A Figura 2 ilustra geometricamente a distância de quarteirão entre dois pontos \mathbf{X}_i e \mathbf{X}_k no \mathbb{R}^2 .

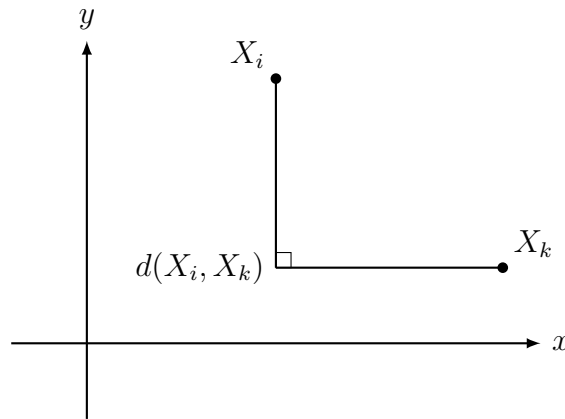


Figura 2: Distância de quarteirão entre os pontos \mathbf{X}_i e \mathbf{X}_k .

3.2.3 Distância de Mahalanobis

Seja $\mathbf{x} = (X_1, \dots, X_p)^T$ vetor aleatório com $\mathbb{E}(\mathbf{x}) = \mu = (\mu_1, \dots, \mu_p)^T$, e $Cov(\mathbf{x}) = \Sigma$. Artes e Barroso (2023) definem a distância de Mahalanobis ao quadrado entre \mathbf{x} e μ como:

$$D_M^2(\mathbf{x}, \mu) = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \quad (3.2.4)$$

Seja $i = 1, \dots, n$ unidades amostrais de dimensão p . Logo, o vetor $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ 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 ao centróide da amostra será

$$D_M^2(\mathbf{x}_i, \bar{\mathbf{x}}) = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (3.2.5)$$

Pela definição trazida na Equação (3.2.4), necessitaríamos saber os valores μ e Σ . Artes e Barroso (2023) argumentam que, em termos práticos, isto não é possível. Portanto, sugerem utilizar os estimadores usuais $\bar{\mathbf{x}}$ e \mathbf{S} , respectivamente, dando suporte para a definição 3.2.5. Os resultados passam a ser aproximados, com melhor desempenho para grandes amostras.

3.2.4 Distância de cossenos

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

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}| |\mathbf{B}| \cos \theta \quad (3.2.6)$$

Ricardo e Berthier (2011) argumentam 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. Da Equação 3.2.6, temos que esta similaridade será:

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

A Figura 3, ilustra o ângulo θ formado entre os vetores \mathbf{A} e \mathbf{B} .

Se considerarmos \mathbf{A} e \mathbf{B} os vetores numéricos (vetorização) das petições de dois

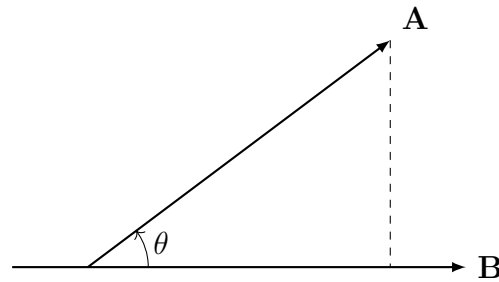


Figura 3: O cosseno como medida de similaridade entre **A** e **B**

processos, conseguimos assim calcular uma medida de similaridade entre eles.

Note que a distância do cosseno não respeita a desigualdade triangular definida em 3.2, portanto não pode ser considerada uma medida de distância estatística naquelas definições.

3.3 Métodos de agrupamento hierárquicos

Segundo von Borries e Wang (2009), métodos hierárquicos buscam dividir ou agrupar um conjunto de dados em uma sequência de partições aninhadas, segundo alguma medida de similaridade (ou dissimilaridade). Estes dividem-se em dois grupos: métodos aglomerativos e divisivos.

3.3.1 Aglomerativo

Este método se inicia com uma observação em cada grupo (passo 0), 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 soma de quadrados das distâncias de todas as possíveis combinações de pares.

A Figura 4 ilustra o fluxograma do algoritmo aglomerativo de forma geral.

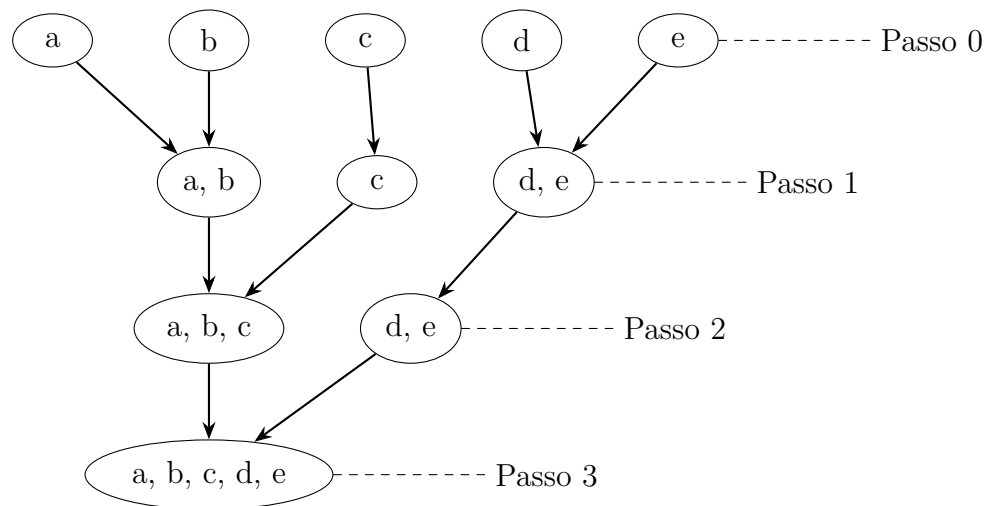


Figura 4: Fluxograma do algoritmo aglomerativo

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. Este método foi primeiramente apresentado por Kaufman e Rousseeuw (1990) com o algoritmo DIANA (*Divisive Analysis* - Análise divisiva).

A Figura 5 ilustra o fluxograma do algoritmo divisivo de forma geral.

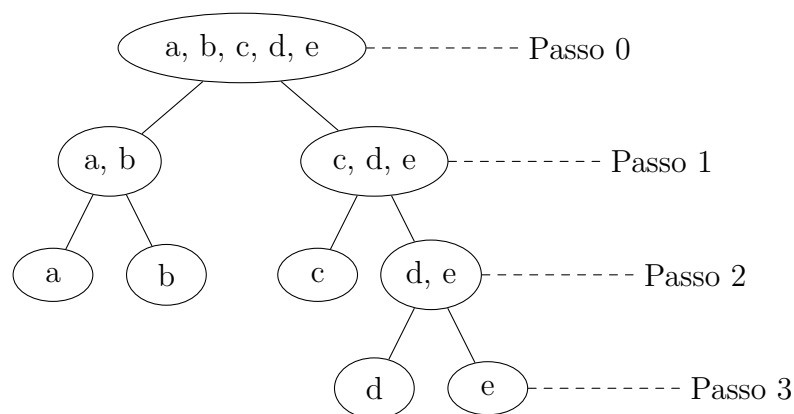


Figura 5: Fluxograma do algoritmo divisivo

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), conforme mostrado em MacQueen et al. (1967)

3.4.1 k-médias

k-médias, geralmente evocado como *k-means*, ou ainda k-medóides (Artes e 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 os pontos que o compõem (Morettin e Singer, 2021).

Seja k o número de grupos em que se deseja agrupar os dados. Seja $\mathbf{x} = \mathbf{x}_1, \dots, \mathbf{x}_n$ o vetor de amostra. O procedimento realizado pelo algoritmo pode ser descrito como:

- Fixar um número de k grupos (O usuário precisa especificar este número);
- Alocar os elementos aleatoriamente aos k grupos, e calcular a medóide $\bar{\mathbf{x}}_k$ de cada grupo;
- Calcular a distância Euclidiana de cada item ao medóide do grupo em que está alocado:

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

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

- Realocar cada item \mathbf{x}_i ao grupo com medóide mais próximo tal que reduza d_{i,\bar{x}_k} . Então, calcular 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 deste trabalho é dado pela petição inicial (texto em formato PDF) dos processos de controle concentrados tramitados ou em tramitação entre 1º de

Janeiro de 2016 e 10 de Abril de 2024. 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 fornecidos pela Secretaria de Tecnologia e Inovação (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 compõem o acervo na data fixada de cada teste.

3.6 Visualização de dados

Uma parte crucial deste trabalho será mostrar de forma simplificada as indicações de agrupamentos fornecidos pela modelagem. Para isso, iremos explorar algumas técnicas de visualização de dados para estes resultados, afim de auxiliar no processo de tomada de decisões. Duas ferramentas particularmente úteis para este tipo de modelagem são dendogramas e *t-SNE*, que serão definidos a seguir.

3.6.1 Dendograma

Pela definição contida em Everitt e 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. A Figura 6 ilustra um dendrograma genérico.

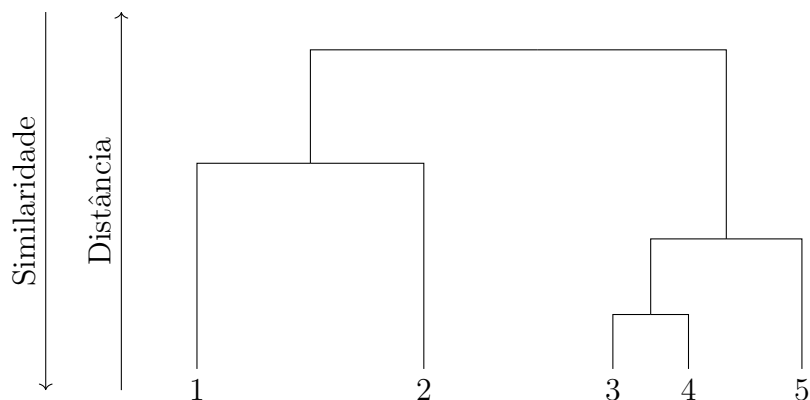


Figura 6: Exemplo simples de um dendograma

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.

A Figura 7 traz o exemplo da técnica *t-SNE* para visualizar o conjunto de dados MNIST, de LeCun et al. (1998)

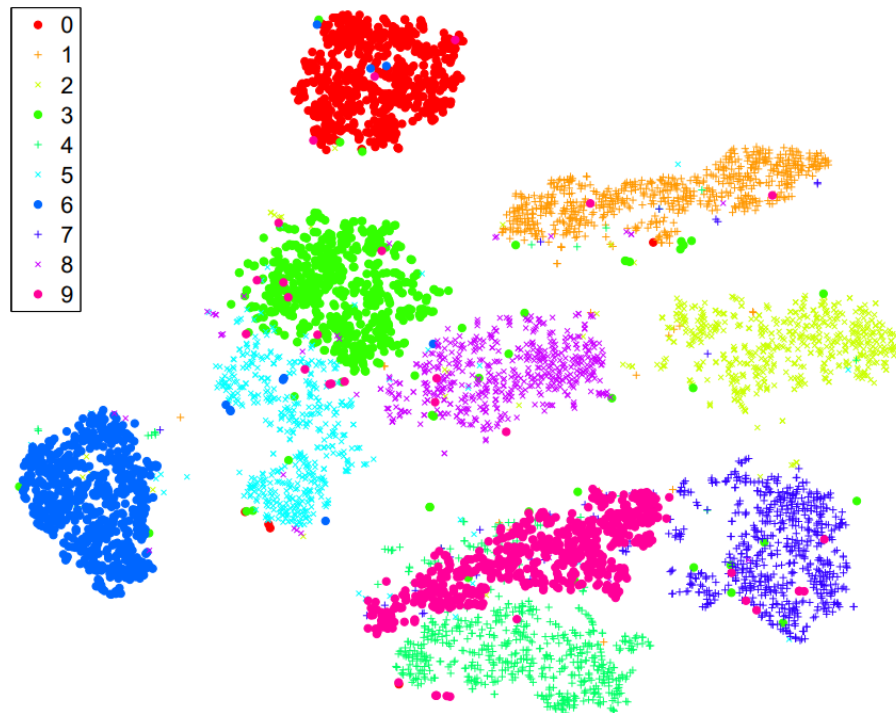


Figura 7: O algoritmo t-SNE aplicado aos dados MNIST.
Fonte: Maaten e Hinton (2008)

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 são:

Tabela 1: Cronograma TCC 1

Atividades	1/2024				
	Mar	Abr	Mai	Jun	Jul
Escolha do tema a ser abordado.					
Levantamento de bibliografia relacionada ao tema.					
Definição do recorte temporal com a AAJ do STF.					
Solicitação dos dados para a STI do STF.					
Solicitação dos algoritmos à STI do STF.					
Revisão de literatura.					
Desenvolvimento da proposta de projeto.					
Análise preliminar do banco de dados.					
Entrega da proposta de projeto.					
Elaboração da apresentação da proposta.					
Manipulação do banco de dados.					
Análise do banco de dados.					
Elaboração do relatório parcial.					
Entrega do relatório parcial a Professora Orientadora.					
Correção do relatório parcial.					
Entrega do relatório parcial a banca.					

Tabela 2: Cronograma TCC 2

Atividades	2/2024				
	Ago	Set	Out	Nov	Dez
Desenvolvimento do modelo e da aplicação.					
Elaboração do relatório final.					
Entrega do relatório final ao Professor Orientador(a).					
Correção do do relatório final.					
Entrega do relatório final para a banca.					

Referências

- ARTES, R.; BARROSO, L. P. *Métodos multivariados de análise estatística*. [S.l.]: São Paulo: Blucher, 2023.
- EVERITT, B.; SKRONDAL, A. *The cambridge dictionary of statistics*. [S.l.]: Cambridge University Press, 2010. v. 4.
- FREITAS, L. J. G. et al. 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. et al. 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.
- MACQUEEN, J. et al. *Some methods for classification and analysis of multivariate observations*. [S.l.], 1967. v. 1. 281–297 p.
- 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.
- von Borries, G.; 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.