Evaluating the Impact of Textual Attributes on Distance-Based Learning from Heterogeneous Data Using Text Embeddings and PCA

Bruno O. Sales[1], **Mirlei M. Silva**[1], **Robespierre Pita**[1]

[1] Institute of Computing – Federal University of Bahia (UFBA)

{mirlei.moura,sales.bruno,robespierre.pita}@ufba.br

Abstract. *Datasets frequently combine numerical, categorical, and textual attributes; however, approaches for mixed data focus primarily on the combination of numerical and categorical attributes. This work investigates the impact of incorporating textual attributes into structured datasets for distance-based machine learning tasks. We propose a hybrid pipeline that represents text using embeddings and PCA and integrates these representations with the remaining attributes. The approach was evaluated on three datasets using clustering and classification algorithms. The results show that the impact of textual attributes varies across datasets and algorithms. In classification, their inclusion yielded significant improvements in five out of six comparisons with numerical and categorical attributes, with the most notable result observed for the E-Commerce dataset, where the KNN F1-Macro increased from* 0.349 *to* 0.516 *(*47.9%*).*

Resumo. *Conjuntos de dados frequentemente combinam atributos numéricos, categóricos e textuais, porém abordagens para dados mistos concentram-se principalmente na combinação de atributos numéricos e categóricos. Este trabalho investiga o impacto da incorporação de atributos textuais em bases de dados estruturadas para tarefas de aprendizado de máquina baseadas em distância. Para isso, propõe-se um pipeline híbrido que representa textos por embeddings e PCA e os integra aos demais atributos. A abordagem foi avaliada em três bases de dados, utilizando algoritmos de agrupamento e classificação. Os resultados mostram que o impacto dos atributos textuais varia conforme a base e o algoritmo. Na classificação, sua inclusão apresentou ganhos significativos em cinco das seis comparações com atributos numéricos e categóricos, destacando-se a base E-Commerce, na qual o F1-Macro do KNN aumentou de* $0,349$ *para* $0,516$ *(*$47,9\%$*).*

1. Introdução

A tipologia de valores armazenados em conjuntos de dados modernos pode ser organizada segundo diferentes perspectivas, como a escala de mensuração [Agresti 2013], a orientação temporal ou espacial [Witten and Frank 2002] e a modalidade de representação [Norris and Maier 2014]. Essas tipologias, amplamente discutidas na literatura, buscam definir tanto as formas de representação dos dados em sistemas de informação quanto as operações matemáticas adequadas para sua análise. Assim, compreender as propriedades dessas representações e suas limitações é fundamental para garantir processos robustos de extração de conhecimento.

Atualmente, conjuntos de dados frequentemente combinam atributos de diferentes tipos e modalidades. Entretanto, a representação conjunta desses atributos em tarefas de aprendizado de máquina permanece um desafio, especialmente em métodos baseados em medidas de distância, que geralmente pressupõem dados homogêneos [Lahat et al. 2015]. Embora a literatura apresente diferentes estratégias para lidar com dados heterogêneos, muitas dessas abordagens consideram apenas combinações de atributos numéricos e categóricos [K. Balaji 2018, Van de Velden et al. 2019], dificultando a incorporação direta da informação semântica presente em dados textuais.

Deste modo, este trabalho investiga se a incorporação da informação textual pode produzir matrizes de distância mais informativas para tarefas de aprendizado de máquina. A hipótese investigada é que a integração de representações semânticas dos atributos textuais aos atributos numéricos e categóricos melhora o desempenho em tarefas de agrupamento e classificação em relação ao uso apenas dos atributos numéricos e categóricos. Para testar essa hipótese, propõe-se um *pipeline* híbrido que representa os atributos textuais por meio de *embeddings*, reduz sua dimensionalidade utilizando *Principal Component Analysis* (PCA) e os integra aos demais atributos (numéricos e categóricos) para o cálculo de matrizes de distância. O impacto dessa integração é avaliado por meio da comparação de diferentes combinações de atributos em cenários de agrupamento e classificação.

As principais contribuições deste trabalho são: (i) a proposição de um *pipeline* para integração de atributos numéricos, categóricos e textuais; (ii) a utilização de *embeddings* e PCA para incorporar informações textuais em matrizes de distância; e (iii) uma avaliação experimental do impacto dessas representações em tarefas de agrupamento e classificação.

2. Trabalhos Relacionados

Um dos trabalhos pioneiros sobre o agrupamento de dados mistos foi realizado por [Huang 1998], no qual foi proposta uma solução para encontrar grupos em bases de dados que contêm variáveis numéricas e categóricas. O trabalho apresenta uma técnica de agrupamento de dados mistos utilizando o algoritmo *k-prototypes*, que permite ao usuário atribuir um peso γ às variáveis categóricas, com o objetivo de balancear a base de dados. Os resultados mostraram menor acurácia para valores baixos de γ; contudo, sua definição adequada depende de conhecimento prévio sobre a importância relativa dos atributos.

No geral, as iniciativas de extração de conhecimento em bases de dados com variáveis mistas são limitadas ao tratamento de variáveis categóricas e numéricas [Wang and Mi 2025, Guo et al. 2025, Van de Velden et al. 2019, Dinh et al. 2021]. Por outro lado, quando variáveis textuais estão disponíveis, pesquisadores, analistas e especialistas em processamento de linguagem natural (PLN) frequentemente aplicam modelos que lidam apenas com *corpus* desta única modalidade. Como exemplo, os trabalhos [Mehta et al. 2021] e [Isler 2025] utilizaram bases que possuíam variáveis numéricas e/ou categóricas, porém, o foco de estudo foi apenas nas variáveis textuais. Essas análises, restritas a um único tipo de variável, impõem limitações à descoberta de conhecimento, uma vez que o uso de todas as informações disponíveis pode sustentar achados mais robustos [Krones et al. 2025]. Esta lacuna de pesquisa constitui a principal motivação deste trabalho, especialmente diante da crescente popularidade dos modelos de linguagem e da disponibilidade de bases de dados com atributos heterogêneos. Nesse contexto, torna-se

necessário desenvolver métodos robustos e confiáveis para extrair conhecimento dessas diferentes fontes de informação.

3. Metodologia

Esta seção apresenta a metodologia utilizada para avaliar o impacto da inclusão de atributos textuais em tarefas de aprendizado de máquina. A seguir, são descritos os métodos de agrupamento e classificação, o cálculo de distância em bases de dados mistas e as bases de dados utilizadas nos experimentos.

3.1. Métodos de agrupamento e classificação em dados mistos

Considere um conjunto de dados $D_{N\times p}$, onde N e p se referem aos números de observações e atributos, respectivamente. Nesta matriz, o $j^{ésimo}$ atributo da $i^{ésima}$ observação é um ponto de dado representado por x_{ij}. Portanto, um registro é um vetor $\mathbf{x}_{i.} = (x_{i1}\ x_{i2}\ \ldots\ x_{ip})_{1\times p}$ contendo p valores da $i^{ésima}$ observação e $\mathbf{x}_{.j} = (x_{1j}\ x_{2j}\ \ldots\ x_{nj})_{1\times n}$ corresponde aos n valores do $j^{ésimo}$ atributo.

Algoritmos de agrupamento *crisp* abrangem abordagens particionais, hierárquicas, estatísticas ou habilitadas por redes neurais [Witten and Frank 2002]. O principal objetivo destes algoritmos é produzir k grupos não sobrepostos, endereçando a complexidade inerente à tarefa de particionar um conjunto de dados dentre k^N possibilidades. Em contraste, algoritmos de classificação assumem a existência de um conjunto de treinamento no qual cada observação $\mathbf{x}_{i.}$ está associada a um rótulo de classe y_i [Zhang et al. 2017]. Nesse contexto, o objetivo é aprender uma função de decisão capaz de predizer o rótulo de classe de novas observações não rotuladas.

Neste trabalho, optamos por implementações e modelos capazes de utilizar matrizes de distância, que permitem representar as relações de proximidade entre observações descritas por atributos heterogêneos. Essas relações são armazenadas em uma estrutura de dados $N \times N$, denominada matriz de distância, que expressa a distância entre cada par de observações $\mathbf{x}_{i.} \in D$. Deste modo, utilizamos os algoritmos *K-medoids* e *Agglomerative* para avaliação de agrupamento, bem como *k-Nearest Neighbors* (KNN) e *Support Vector Machines* (SVM) para classificação. Todos os métodos foram implementados utilizando a biblioteca Scikit-learn [Pedregosa et al. 2011] e suas extensões, empregando matrizes de distância por meio do parâmetro *precomputed*.

O algoritmo Kmedoids [Kaufman and Rousseeuw 2009] objetiva minimizar a distância entre os pontos de um *cluster* e o elemento central que o representa. A principal característica desse método de particionamento reside na escolha de observações reais $\mathbf{x}_{i.} \in D$, denominadas medóides, como representantes dos *clusters*. Alternativamente, métodos hierárquicos Aglomerativos constroem uma estrutura de agrupamento por meio de sucessivas fusões entre clusters. Inicialmente, cada observação $\mathbf{x}_{i.}$ é considerada um cluster individual. A cada iteração, os dois clusters mais próximos são combinados de acordo com um critério de ligação (*linkage*), produzindo uma hierarquia de partições representada por um dendrograma. Esse processo é repetido até que todos os elementos pertençam a um único *cluster* ou até que se obtenha um número desejado de grupos K.

O KNN classifica uma observação $\mathbf{x}_{i.}$ a partir da classe majoritária entre seus K vizinhos mais próximos no conjunto de treinamento. A proximidade entre as observações é determinada por uma medida de distância definida no espaço de atributos, permitindo

que o método utilize diretamente representações derivadas das matrizes de distância obtidas no processo de integração de dados [Cover and Hart 1967]. Por sua vez, o SVM busca construir uma superfície de decisão que maximize a margem de separação entre classes distintas [Cortes and Vapnik 1995].

3.2. Calculando distância em dados mistos

Neste trabalho, para agrupar os atributos numéricos e categóricos foi utilizada a medida de distância proposta por [Amir Ahmad 2007]. Essa medida é utilizada para lidar com objetos com atributos numéricos e categóricos. A mesma é uma modificação da equação proposta por [Huang 1998]. Enquanto [Huang 1998] sugere que o usuário indique a importância dos atributos, [Amir Ahmad 2007] propõe que a importância dos atributos seja calculada automaticamente com base nos dados. Outra modificação significativa é que a função de distância para atributos categóricos não assume uma medida binária ou discreta. Em vez disso, a distância entre dois valores de atributos distintos é computada como uma função da distribuição geral e da coocorrência desses valores com outros atributos [Ahmad and Dey 2007]. O cálculo de coocorrência é definido por $\vartheta(d_i, C_j) = \sum_{t=1}^{m_r} w_t(d_{it}^r - C_{jt}^r)^2 + \sum_{t=1}^{m_c} \Omega(d_{it}^c, C_{jt}^c)^2$, onde a parte numérica é expressa por $\sum_{t=1}^{m_r} w_t(d_{it}^r - C_{jt}^r)^2$, $(d_{it}^r - C_{jt}^r)^2$ indica a distância do objeto d_{it}^r ao centroide do *cluster*, e w_t, representa a significância do $t^{ésimo}$ atributo numérico. A parte categórica é representada por $\sum_{t=1}^{m_c} \Omega(d_{it}^c, C_{jt}^c)^2$ que expressa a distância do objeto ao centróide do *cluster* em termos de variáveis categóricas, e Ω representa um cálculo que leva em consideração a distribuição de valores categóricos dentro do cluster, em vez de depender de uma simples medida binária ou discreta [Ahmad and Dey 2007].

Para bases contendo variáveis numéricas e categóricas, a matriz de distância foi computada a partir de uma implementação em Python da distância Ahmad, originalmente disponibilizada na biblioteca *kmed* do ambiente R. Para os atributos textuais, o *pipeline* inclui uma etapa de extração de características baseada em *embeddings* contextualizados. Esse processo é descrito a seguir.

3.2.1. Extraindo características numéricas a partir de *Embeddings*

Modelos de *Embeddings* mapeiam objetos não numéricos, tais como texto, imagens e áudio, em um espaço vetorial contínuo. Isso possibilita trabalhar com dados não numéricos em um modelo matemático [Rajvardhan Patil 2023]. No contexto deste trabalho, utilizamos *Embeddings* contextuais, especificamente o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019], que utiliza uma arquitetura de *Transformers* bidirecional pré-treinada em um *corpus* de alta dimensionalidade.

Deste modo, cada atributo textual foi submetido à geração de representações contextualizadas por meio do modelo BERTimbau *bert-base-portuguese-cased* [Fábio Souza 2020]. Como o modelo produz vetores contextualizados para cada token da sequência de entrada, foi aplicada a estratégia de *mean pooling* sobre esses vetores, obtendo-se uma representação densa única para cada registro. Essa abordagem é

amplamente utilizada para a geração de representações semânticas de sentenças baseadas em BERT [Reimers and Gurevych 2019].

Devido à alta dimensionalidade das representações produzidas pelo BERT, aplicou-se a técnica de Análise de Componentes Principais (Principal Component Analysis – PCA) para redução de dimensionalidade. O número de componentes principais $Q = PCA_1, PCA_2, \ldots, PCA_m$ foi definido para preservar pelo menos 85% da variância explicada, equilibrando preservação da informação e viabilidade computacional, uma vez que o limiar de 95% resultou em alta dimensionalidade para o cálculo das matrizes de distância. As componentes principais resultantes foram então incorporadas como novos atributos numéricos em cada registro $\mathbf{x}_{i.}$. Desta forma, o número de atributos em $\mathbf{x}_{i.}$ passa a ser $p + |Q|$. Uma vez criadas as colunas numéricas obtidas pelo PCA, torna-se possível a utilização da distância Ahmad para calcular a distância dos dados mistos (numéricos e categóricos) resultantes.

3.3. Bases de dados

Foram utilizadas três bases de dados de diferentes domínios, contendo atributos numéricos, categóricos e textuais. A Tabela 1 apresenta suas principais características, incluindo a quantidade de atributos por tipo, a distribuição das observações numéricas, o número de categorias, a contagem de termos das variáveis textuais e a distribuição das variáveis-alvo utilizadas na classificação. A seguir, são descritos a origem dos dados, os atributos textuais e os rótulos considerados em cada base.

3.3.1. BRATECA (*Brazilian Tertiary Care Dataset*)

BRATECA é um conjunto de dados clínicos composto por registros de internações provenientes de dez hospitais localizados em dois estados brasileiros. A base apresenta informações sobre descritores de pacientes, prescrições médicas, exames laboratoriais e notas clínicas em texto livre [Dias and Ulbrich 2022] [1]. Neste trabalho foram utilizadas as tabelas de admissão, prescrição e itens da prescrição, as quais foram integradas e filtradas para incluir apenas os pacientes internados na Unidade de Terapia Intensiva (UTI). O processamento resultou em um conjunto de dados composto por atributos numéricos, categóricos e textuais. As variáveis textuais incorporam informações semânticas sobre os regimes terapêuticos adotados, enquanto a variável-alvo representa o desfecho da internação, classificado como alta clínica ou óbito.

3.3.2. Kaggle Brazilian E-Commerce Public Dataset by Olist (KaggleBE)

Este conjunto de dados refere-se a transações de comércio eletrônico realizadas na Olist, um marketplace brasileiro [2]. A base contém informações de aproximadamente 100 mil pedidos realizados entre 2016 e 2018, abrangendo aspectos como produtos, pagamentos, logística e avaliações dos consumidores. Neste estudo, foram integradas as tabelas *olist-order-reviews-dataset*, *olist-products-dataset* e *olist-order-items-dataset*. Os comentários

[1] https://physionet.org/content/brateca/1.0/

[2] https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

das avaliações (*review comments*) constituem os atributos textuais, enquanto a nota atribuída pelo consumidor (*review score*), variando de 1 a 5, é utilizada como variável-alvo. Diferentemente das demais bases, nas quais uma ou duas modalidades predominam, esta base apresenta representação mais homogênea dos diferentes tipos de atributos, contemplando quantidades semelhantes de variáveis numéricas, categóricas e textuais.

3.3.3. Cursos da Escola Nacional de Administração Pública (ENAP)

Esta base de dados foi produzida a partir de uma amostra de 5% dos cursos oferecidos pela ENAP em 2025, estratificados por mês de início. O conjunto de registros contém atributos que descrevem cursos, participantes e profissionais na plataforma da Escola Virtual de Governo (E.V.G), provendo informações que podem subsidiar diversas análises demográficas. Os atributos textuais correspondem ao nome da turma, nome do curso, temática e conteudista, enquanto a situação da turma é utilizada como variável-alvo, composta pelas classes *CONFIRMADA*, *CONCLUIDA* e *CRIADA*. A base apresenta distribuição desbalanceada entre essas classes, com predominância de *CONFIRMADA*, permitindo avaliar a incorporação de atributos heterogêneos nesse cenário.

Tabela 1. Resumo das bases de dados utilizadas. Variáveis categóricas (CAT) indicam número de categorias distintas; textuais (TXT) indicam número de termos distintos e tamanho médio de texto; RÓTULO indica a distribuição da variável-alvo de cada base.

Base de dados	Tipo	Qtd. atributos	Resumo estatístico
BRATECA N = 4.217	NUM	28	média: 12,19, desvio padrão: 14,14; min: 0, max: 5.171
	CAT	7	2 a 5 categorias
	TXT	3	3.965 termos distintos; tam. médio: 57,61 palavras
	RÓTULO	1	ALTA_CLINICA: 75,0% (N=3.163) OBITO: 25,0% (N=1.054)
ECOMMERCE N = 10.516	NUM	2	média: 72,09, desvio padrão: 74,10; min: 0, max: 999,99
	CAT	2	65 e 13 categorias
	TXT	2	7.607 termos distintos; tam. médio: 6,99 palavras
	RÓTULO	1	1: 19,8% (N=2.086) 2: 5,2% (N=545) 3: 8,1% (N=852) 4: 14,7% (N=1.549) 5: 52,1% (N=5.484)
ENAP N = 6.602	NUM	16	média: 13,53, desvio padrão: 18,96; min: 0, max: 1.304
	CAT	2	27 e 10 categorias
	TXT	4	1.448 termos distintos; tam. médio: 30,31 caracteres
	RÓTULO	1	CONFIRMADA: 89,4% (N=5.904) CONCLUIDA: 9,6% (N=637) CRIADA: 0,9% (N=61)

3.4. Métricas de validação

Para os algoritmos de agrupamento, utilizou-se o coeficiente de Silhueta, definido por $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$, em que $a(i)$ corresponde à distância média entre o ponto i e os demais elementos do mesmo grupo, enquanto $b(i)$ representa a menor distância média entre i e os elementos de outros grupos. Os valores variam entre -1 e 1, indicando, respectivamente, má atribuição e boa separação dos *clusters* [Rousseeuw 1987].

Para classificação, foram utilizadas as métricas de Acurácia, definida como $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, e F1-Macro, definida por $F1_{Macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i$, onde C representa o número de classes. Por atribuir o mesmo peso a todas as clas-

ses, o F1-Macro é particularmente adequado para conjuntos de dados desbalanceados [Grandini et al. 2020].

4. Avaliação experimental

A Figura 1 ilustra os quatro passos que compõem o fluxo experimental: seleção das variáveis, cálculo das matrizes de distância, execução dos modelos e avaliação. São consideradas três combinações de atributos: somente numéricos (*num*), numéricos e categóricos (*num+cat*) e todos os atributos disponíveis (*num+cat+txt*). Para cada combinação, é construída uma matriz de distância específica. A distância Euclidiana é aplicada aos atributos numéricos, enquanto a distância de Ahmad é utilizada para a combinação de atributos numéricos e categóricos. Para incorporar os atributos textuais, estes são inicialmente representados por meio de modelos de *embedding* pré-treinados e, em seguida, submetidos à redução de dimensionalidade por PCA. As representações resultantes são combinadas com os atributos numéricos e categóricos originais, e a distância de Ahmad é então utilizada para gerar a matriz de distância considerando todos os atributos do conjunto de dados.

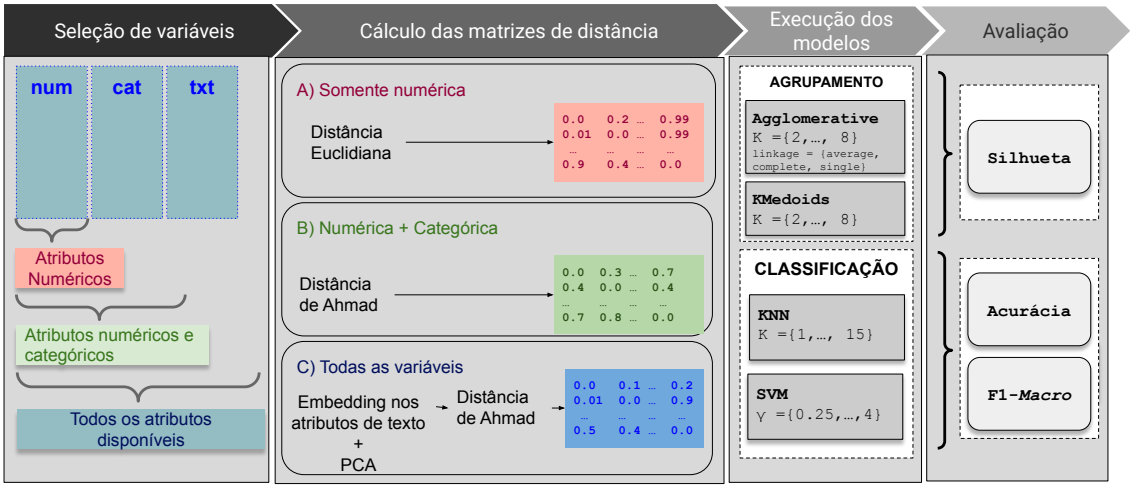


Figura 1. Desenho do experimento.

Para possibilitar a análise estatística da variabilidade dos resultados, os algoritmos envolvendo aleatoriedade foram repetidos com 150 sementes distintas. No agrupamento, K-Medoids e Agglomerative foram avaliados para $K \in 2, \ldots, 8$; para o K-Medoids, as 150 execuções independentes para cada K variam a inicialização dos medóides, enquanto o Agglomerative é determinístico.

Na classificação, o KNN foi avaliado para valores ímpares de K entre 1 e 15, e o SVM para $\gamma = 0{,}25, 0{,}5, 1, 2, 4 \times \gamma_0$, sendo γ_0 estimado pela heurística da mediana. A avaliação utilizou *Stratified 10-Fold Cross-Validation*, repetida com 150 sementes distintas, variando as partições dos dados.

5. Resultados e discussão

As Tabelas 2 e 3 apresentam os resultados para as tarefas de agrupamento, onde BRA refere-se à base BRATECA, ECO à base E-Commerce KaggleBE e ENP à base ENAP. Observa-se que a inclusão de atributos categóricos apresentou ganhos nos valores de silhueta para todos os conjuntos de dados e algoritmos avaliados. Em contrapartida, a adi-

ção dos atributos textuais não resultou em melhorias quando comparada à representação num+cat, sendo observadas reduções estatisticamente significativas em diversos cenários.

Uma possível explicação para esse comportamento consiste no aumento da dimensionalidade da representação final. Embora os embeddings textuais tenham sido reduzidos por meio de PCA, a preservação de 85% da variância resultou na geração de 5 a 107 atributos numéricos para cada atributo textual. Em cenários de alta dimensionalidade, as medidas de distância tendem a se tornar menos discriminativas, dificultando a identificação de grupos compactos e bem separados, fenômeno amplamente discutido na literatura de agrupamento em espaços de alta dimensão [Parsons et al. 2004].

Tabela 2. Silhueta média $\pm$ desvio-padrão por representação.

Algoritmo	Dataset	Representação	Silhueta
K-medoids	BRA	num	0,201 $\pm$ 0,082
K-medoids	BRA	num+cat	**0,338 $\pm$ 0,137**
K-medoids	BRA	num+cat+txt	0,336 $\pm$ 0,136
K-medoids	ECO	num	0,502 $\pm$ 0,038
K-medoids	ECO	num+cat	**0,808 $\pm$ 0,016**
K-medoids	ECO	num+cat+txt	0,757 $\pm$ 0,029
K-medoids	ENP	num	0,417 $\pm$ 0,098
K-medoids	ENP	num+cat	**0,771 $\pm$ 0,055**
K-medoids	ENP	num+cat+txt	0,768 $\pm$ 0,055
Agglomerative	BRA	num	0,732 $\pm$ 0,082
Agglomerative	BRA	num+cat	**0,933 $\pm$ 0,014**
Agglomerative	BRA	num+cat+txt	0,932 $\pm$ 0,014
Agglomerative	ECO	num	0,842 $\pm$ 0,040
Agglomerative	ECO	num+cat	0,945 $\pm$ 0,022
Agglomerative	ECO	num+cat+txt	**0,945 $\pm$ 0,012**
Agglomerative	ENP	num	0,941 $\pm$ 0,000
Agglomerative	ENP	num+cat	**0,992 $\pm$ 0,000**
Agglomerative	ENP	num+cat+txt	0,992 $\pm$ 0,000

Tabela 3. Testes estatisticos para agrupamento (Wilcoxon pareado, correcao de Bonferroni, $\alpha = 0{,}05$).

Dataset	Algoritmo	Comparação	Δ med.	r	p (Bon.)	Sig.	num+cat+txt
BRA	**K-medoids**	**num**	**0,06144**	**1,000**	**<0,001**	**sim**	**superior**
BRA	K-medoids	num+cat	-0,00104	-0,246	0,0179	sim	inferior
ECO	K-medoids	num	-0,01431	-0,308	0,0021	sim	inferior
ECO	K-medoids	num+cat	-0,27846	-1,000	<0,001	sim	inferior
ENP	**K-medoids**	**num**	**0,10197**	**0,998**	**<0,001**	**sim**	**superior**
ENP	K-medoids	num+cat	-0,05245	-0,964	<0,001	sim	inferior
BRA	**Agglomerative**	**num**	**0,21063**	**1,000**	**<0,001**	**sim**	**superior**
BRA	Agglomerative	num+cat	-0,00020	-0,186	0,9097	nao	n.s.
ECO	**Agglomerative**	**num**	**0,08823**	**1,000**	**<0,001**	**sim**	**superior**
ECO	Agglomerative	num+cat	-0,03894	-0,965	<0,001	sim	inferior
ENP	**Agglomerative**	**num**	**0,07648**	**1,000**	**<0,001**	**sim**	**superior**
ENP	Agglomerative	num+cat	-0,00057	-1,000	<0,001	sim	inferior

n.s. = nao significativo.

As Tabelas 4, 5, 6 e 7 apresentam os resultados obtidos para as tarefas de classificação. A significância estatística foi avaliada por meio do teste de Wilcoxon pareado com correção de Bonferroni ($\alpha = 0,05$). Os maiores ganhos foram observados na base E-Commerce KaggleBE, na qual a representação num+cat+txt apresentou melhorias estatisticamente significativas em relação às representações num e num+cat para os algoritmos KNN e SVM. Em comparação à representação num+cat, os ganhos em F1-Macro

alcançaram $47,9\%$ para o KNN e $164,1\%$ para o SVM. Para a acurácia, os ganhos foram de $40,6\%$ e $155,2\%$, respectivamente. Considerando que essa base possui apenas dois atributos numéricos e dois categóricos, os resultados sugerem que as informações textuais carregam parte relevante do poder discriminativo necessário para distinguir as classes. Além disso, trata-se da única base multiclasse com distribuição relativamente equilibrada entre as classes, o que pode favorecer o aproveitamento das informações semânticas contidas nos atributos textuais.

As bases BRATECA e ENAP também apresentaram benefícios decorrentes da incorporação dos atributos textuais em diversos cenários. Em BRATECA, observou-se melhora consistente tanto para Acurácia quanto para F1-Macro. Já em ENAP, embora tenham sido observados ganhos para alguns algoritmos e métricas, os efeitos foram mais modestos e o KNN apresentou desempenho ligeiramente inferior após a inclusão dos atributos textuais. Esse comportamento pode estar relacionado ao forte desbalanceamento da variável-alvo.

Diferentemente do observado nos experimentos de agrupamento, os atributos textuais mostraram-se particularmente úteis em tarefas supervisionadas. Enquanto os agrupamentos foram predominantemente influenciados pelos atributos categóricos, a incorporação do texto contribuiu para aumentar a capacidade discriminativa dos classificadores, sobretudo em cenários com menor disponibilidade de atributos estruturados. Esses resultados sugerem que as representações textuais capturam informações semânticas relevantes para a separação das classes, mesmo quando tais informações não se refletem diretamente na formação de grupos mais compactos e bem separados.

Tabela 4. Acurácia média $\pm$ desvio-padrão por representação.

Algoritmo	Dataset	Representação	Acurácia
KNN	BRA	num	0,837 $\pm$ 0,002
KNN	BRA	num+cat	0,838 $\pm$ 0,002
KNN	BRA	num+cat+txt	**0,839 $\pm$ 0,002**
KNN	ECO	num	0,463 $\pm$ 0,003
KNN	ECO	num+cat	0,463 $\pm$ 0,003
KNN	ECO	num+cat+txt	**0,651 $\pm$ 0,002**
KNN	ENP	num	0,859 $\pm$ 0,002
KNN	ENP	num+cat	0,859 $\pm$ 0,002
KNN	ENP	num+cat+txt	**0,862 $\pm$ 0,001**
SVM	BRA	num	**0,735 $\pm$ 0,094**
SVM	BRA	num+cat	0,661 $\pm$ 0,046
SVM	BRA	num+cat+txt	0,664 $\pm$ 0,046
SVM	ECO	num	0,186 $\pm$ 0,022
SVM	ECO	num+cat	0,145 $\pm$ 0,018
SVM	ECO	num+cat+txt	**0,370 $\pm$ 0,116**
SVM	ENP	num	**0,571 $\pm$ 0,260**
SVM	ENP	num+cat	0,460 $\pm$ 0,203
SVM	ENP	num+cat+txt	0,462 $\pm$ 0,205

6. Conclusão e Trabalhos Futuros

Este trabalho investigou o impacto da incorporação de variáveis textuais na análise de dados estruturados, considerando um *pipeline* híbrido para o processamento de atributos numéricos, categóricos e textuais considerando três conjuntos de dados.

Os resultados experimentais demonstraram que as variáveis categóricas desempenham papel fundamental na estruturação global dos agrupamentos. Na classificação, os experimentos indicaram que a inclusão de informação semântica proveniente dos embeddings textuais contribui para melhorar a capacidade discriminativa dos modelos. Esses

Tabela 5. F1-Macro média ± desvio-padrão por representação.

Algoritmo	Dataset	Representação	F1-Macro
KNN	BRA	num	0,758 ± 0,003
KNN	BRA	num+cat	0,759 ± 0,003
KNN	BRA	num+cat+txt	**0,761 ± 0,003**
KNN	ECO	num	0,349 ± 0,003
KNN	ECO	num+cat	0,349 ± 0,003
KNN	ECO	num+cat+txt	**0,516 ± 0,003**
KNN	ENP	num	0,400 ± 0,006
KNN	ENP	num+cat	0,400 ± 0,006
KNN	ENP	num+cat+txt	**0,402 ± 0,006**
SVM	BRA	num	**0,674 ± 0,131**
SVM	BRA	num+cat	0,600 ± 0,091
SVM	BRA	num+cat+txt	0,603 ± 0,092
SVM	ECO	num	0,127 ± 0,051
SVM	ECO	num+cat	0,103 ± 0,027
SVM	ECO	num+cat+txt	**0,272 ± 0,112**
SVM	ENP	num	**0,325 ± 0,137**
SVM	ENP	num+cat	0,260 ± 0,099
SVM	ENP	num+cat+txt	0,262 ± 0,100

Tabela 6. Testes estatísticos para classificação – Acurácia (Wilcoxon pareado, correção de Bonferroni, $\alpha = 0{,}05$).

Dataset	Algoritmo	Comparação	Δ med.	r	p (Bon.)	Sig.	num+cat+txt
BRA	**KNN**	**num**	**0,00110**	**1,000**	**<0,001**	**sim**	**superior**
BRA	**KNN**	**num+cat**	**0,00077**	**1,000**	**<0,001**	**sim**	**superior**
ECO	**KNN**	**num**	**0,17759**	**1,000**	**<0,001**	**sim**	**superior**
ECO	**KNN**	**num+cat**	**0,17766**	**1,000**	**<0,001**	**sim**	**superior**
ENP	**KNN**	**num**	**0,00109**	**0,999**	**<0,001**	**sim**	**superior**
ENP	**KNN**	**num+cat**	**0,00108**	**0,999**	**<0,001**	**sim**	**superior**
BRA	SVM	num	-0,09201	-1,000	<0,001	sim	inferior
BRA	**SVM**	**num+cat**	**0,00126**	**0,998**	**<0,001**	**sim**	**superior**
ECO	**SVM**	**num**	**0,13481**	**1,000**	**<0,001**	**sim**	**superior**
ECO	**SVM**	**num+cat**	**0,15847**	**1,000**	**<0,001**	**sim**	**superior**
ENP	SVM	num	-0,10718	-1,000	<0,001	sim	inferior
ENP	**SVM**	**num+cat**	**0,00646**	**0,998**	**<0,001**	**sim**	**superior**

Tabela 7. Testes estatísticos para classificação – F1-Macro (Wilcoxon pareado, correção de Bonferroni, $\alpha = 0{,}05$).

Dataset	Algoritmo	Comparação	Δ med.	r	p (Bon.)	Sig.	num+cat+txt
BRA	**KNN**	**num**	**0,00151**	**1,000**	**<0,001**	**sim**	**superior**
BRA	**KNN**	**num+cat**	**0,00102**	**0,999**	**<0,001**	**sim**	**superior**
ECO	**KNN**	**num**	**0,14503**	**1,000**	**<0,001**	**sim**	**superior**
ECO	**KNN**	**num+cat**	**0,14506**	**1,000**	**<0,001**	**sim**	**superior**
ENP	KNN	num	-0,00140	-0,932	<0,001	sim	inferior
ENP	KNN	num+cat	-0,00143	-0,931	<0,001	sim	inferior
BRA	SVM	num	-0,09475	-1,000	<0,001	sim	inferior
BRA	**SVM**	**num+cat**	**0,00135**	**0,999**	**<0,001**	**sim**	**superior**
ECO	**SVM**	**num**	**0,10050**	**1,000**	**<0,001**	**sim**	**superior**
ECO	**SVM**	**num+cat**	**0,11792**	**1,000**	**<0,001**	**sim**	**superior**
ENP	SVM	num	-0,05555	-1,000	<0,001	sim	inferior
ENP	**SVM**	**num+cat**	**0,00351**	**1,000**	**<0,001**	**sim**	**superior**

achados reforçam que a utilização de representações mistas, integrando diferentes tipos

de dados, constitui uma estratégia promissora para ampliar a expressividade das representações utilizadas em tarefas de aprendizado de máquina.

Dessa forma, a principal contribuição deste trabalho consiste em demonstrar que informações textuais obtidas por meio de modelos de linguagem podem ser integradas de maneira eficiente a atributos estruturados. Como trabalhos futuros, pretende-se investigar estratégias alternativas para integração de atributos textuais em matrizes de distância, incluindo técnicas de redução de dimensionalidade distintas do PCA e métodos de seleção de atributos capazes de preservar informações relevantes para agrupamento.

Referências

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.

Ahmad, A. and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data amp; Knowledge Engineering*, 63(2):503–527.

Amir Ahmad, L. D. (2007). A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*, 28(1):110–118.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dias, H. and Ulbrich, A. H. D. P. d. (2022). BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. *PhysioNet*. Version 1.0.

Dinh, D.-T., Huynh, V.-N., and Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, 571:418–442.

Fábio Souza, Rodrigo Nogueira, R. L. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. *Lecture Notes in Computer Science*, page 403–417.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview.

Guo, C., Tang, H., Niu, B., and Zhong, H. (2025). Clustering mixed data using multi-objective bacterial colony optimization with diverse learning mechanisms. *Swarm and Evolutionary Computation*, 98:102119.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304.

Isler, C. (2025). Words matter: Forecasting economic downside risks with corporate textual data. *arXiv preprint arXiv:2511.04935*.

K. Balaji, K. L. (2018). Clustering algorithms for mixed datasets: a review. *Int J Pure Appl Math*, 18(7):547–56.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Krones, F., Marikkar, U., Parsons, G., Szmul, A., and Mahdi, A. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Mehta, V., Bawa, S., and Singh, J. (2021). Weclustering: word embeddings based text clustering technique for large datasets. *Complex amp; Intelligent Systems*, 7(6):3211–3224.

Norris, S. and Maier, C. D. (2014). *Interactions, images and texts: A reader in multimodality*, volume 11. Walter de Gruyter GmbH & Co KG.

Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: A review. *SIGKDD Explorations*, 6(1):90–105.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rajvardhan Patil, Sorio Boit, V. G. J. N. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Van de Velden, M., Iodice D'Enza, A., and Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3):e1456.

Wang, H. and Mi, J. (2025). Intuitive-k-prototypes: A mixed data clustering algorithm with intuitionistic distribution centroid. *Pattern Recognition*, 158:111062.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Zhang, S., Li, X., Zong, M., Zhu, X., and Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19.