

Aprendizado de Máquina para Estruturação de Texto Livre e Predição de Eventos Cardiovasculares em Dados Clínicos Multicêntricos



Mayara Martins Perroni

Orientador: Prof. Dr. Paulo Mazzoncini de Azevedo Marques

Faculdade de Medicina de Ribeirão Preto

Departamento de Imagens Médicas, Hematologia e Oncologia Clínica

Divisão de Ciências das Imagens e Física Médica



Introdução

- Categorização de Fatores de Risco e Dados em Texto Livre
- Otimização da Categorização e Vetorização
- Aprendizado de Máquina na Saúde

Categorização de Fatores de Risco e Dados em Texto Livre

- Dados de prontuários eletrônicos de saúde contêm descrições detalhadas de sintomas, condições de saúde, e comportamentos de risco que muitas vezes não estão disponíveis em formatos estruturados ou numéricos
- Uso de técnicas avançadas de Processamento de Linguagem Natural (PLN) para transformar esses textos em representações que possam ser interpretadas por modelos preditivos
- O modelo se torna capaz de utilizar informações não estruturadas que seriam negligenciadas em uma abordagem exclusivamente baseada em dados estruturados

Otimização da Categorização e Vetorização

- A categorização é o processo pelo qual dados textuais são classificados em grupos ou categorias que representam fatores de risco ou sintomas relevantes
- A categorização adequada depende da definição de um sistema de rótulos que capture com precisão a variação linguística e as nuances presentes nos textos dos prontuários
- A categorização por si só não é suficiente. Para que essas informações categorizadas sejam úteis em modelos preditivos, elas precisam ser transformadas em vetores numéricos
- Esse processo de vetorização envolve a conversão de textos em representações matemáticas que reflitam a relevância e o contexto dos termos no corpo textual

Vetorização

- Técnicas como o TF-IDF (Term Frequency-Inverse Document Frequency) são amplamente utilizadas para vetorização
- Técnicas de vetorização ajudam a identificar palavras mais informativas ao atribuir pesos proporcionais à sua frequência e relevância nos documentos
- A otimização desse processo implica encontrar o equilíbrio entre a granularidade das categorias e a dimensionalidade dos vetores resultantes
- Essa otimização é focada em evitar tanto a perda de informações quanto a introdução de ruídos

Aprendizado de Máquina na Saúde

- Capacidade de analisar grandes volumes de dados e detectar padrões complexos
- Estudos recentes demonstram que algoritmos de aprendizado supervisionado, como redes neurais e máquinas de vetores de suporte (SVM), são capazes de prever MACE com alta precisão
- Os campos textuais, quando processados e integrados corretamente, podem fornecer insights adicionais sobre o estado de saúde do paciente, com informações que não são capturadas em dados estruturados tradicionais (Polat Erdeniz et al., 2023)

Problema



Doenças cardiovasculares (DCVs) são a principal causa de mortalidade mundial, resultando em milhões de mortes anuais e impondo um pesado ônus econômico sobre os sistemas de saúde.



Os métodos tradicionais de estratificação de risco cardiovascular, como o ESC SCORE e o Framingham Risk Score, ainda se baseiam em um número limitado de fatores de risco e carecem de precisão e personalização.



A validação desses modelos em contextos clínicos reais ainda enfrenta obstáculos, especialmente pela variabilidade dos dados clínicos entre diferentes centros e populações.

Justificativa

- Registros eletrônicos de saúde (RES) possuem grande potencial para análise clínica, mas:

Dados heterogêneos: Estruturas diferentes entre sistemas

Textos livres: Difíceis de padronizar e analisar

Impacto: Dados despadronizados comprometem a qualidade das análises automatizadas

Hipóteses

- **H1:** O uso de abordagens de PLN, como TF-IDF, em conjunto com Modelos de Linguagem de Grande Escala (LLMs), permite a vetorização eficiente de informações não estruturadas em texto livre extraídas de EHRs para uso em modelos convencionais de aprendizado de máquina.
- **H0:** O uso de PLN, TF-IDF e LLMs em informações de EHRs não tem efeito significativo na vetorização e na estruturação de dados textuais para uso em modelos de aprendizado de máquina.

Objetivo Central

- Converter informações textuais livres em dados estruturados, utilizando modelos de linguagem de grande escala (LLMs), de modo a **maximizar** o aproveitamento desses dados e aprimorar a acurácia dos modelos preditivos. A **padronização e organização dos dados** são conduzidas de forma a **atender às necessidades de um contexto multicêntrico e heterogêneo**, contribuindo para a melhoria da análise preditiva e para o **avanço na precisão e abrangência das previsões de risco cardiovascular**.

Objetivo Complementar

- Potencializar a inclusão dessas features adicionais nos modelos de predição do projeto PRE-CARE ML (Prevendo eventos cardiovasculares usando aprendizado de máquina), projeto multicêntrico regular com financiamento da FAPESP(#2021/06137-4)



Materiais e Métodos

Materiais e Métodos

1. Revisão Bibliográfica e Estudo de Procedimentos ETL
2. Implementação de ETL e Criação do Banco de Dados
3. Integração e Estruturação dos Dados
4. Categorização de Dados Não Estruturados
5. Classificação de Dados Textuais Usando Modelos de Aprendizado de Máquina

1. Revisão Bibliográfica e Estudo de Procedimentos ETL

2. Implementação de ETL e Criação do Banco de Dados

2. Implementação de ETL e Criação do Banco de Dados



3. Integração e Estruturação dos Dados

4. Categorização de Dados Não Estruturados

Regra para categorizar os dados

```
UPDATE `fmrp-usp-br.medical_terms.Smoking_Processed`
SET CATEGORY_2 = (
  CASE
    WHEN LOWER(CTU_INFORMACAO) LIKE '%tabagismo%' OR LOWER(CTU_INFORMACAO) LIKE '%tabagista%'
    THEN
      CASE
        WHEN LOWER(CTU_INFORMACAO) LIKE '%nega tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%negam tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%nega etilismo e tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%ex-tabagista%' OR
              LOWER(CTU_INFORMACAO) LIKE '%ex tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%ex- tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%ex - tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%nega etilismo ou tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%( ) tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%nega etilismo, tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%cessou tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%negou tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%cessado tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%negado etilismo ou tabagismo%' OR
              LOWER(CTU_INFORMACAO) LIKE '%passado de tabagismo%'
        THEN 'nosmoke'
        ELSE 'smoke'
      END
    ELSE 'nosmoke'
  END
);
```

Consulta para binarizar categorias

```
SELECT
  COD_PACIENTE,
  DTA_HOR_CADASTRO,
  CTU_INFORMACAO,
  CATEGORY_2
FROM (
  SELECT
    COD_PACIENTE,
    DTA_HOR_CADASTRO,
    CTU_INFORMACAO,
    CATEGORY_2,
    ROW_NUMBER() OVER (PARTITION BY CTU_INFORMACAO ORDER BY CTU_INFORMACAO) AS rn
  FROM `fmrp-usp-br.medical_terms.Smoking_Processed`
  WHERE COD_PACIENTE IS NOT NULL
)
WHERE rn = 1
LIMIT 5000;
```

5. Classificação de Dados Textuais Usando Modelos de Aprendizado de Máquina

Exemplo do target gerado

CTU INFORMACAO	Y
# hipotireoidismo subclínico # tabagismo ativo # abril/2019: herniorrafia	1
# hábitos de vida: tabagismo desde os 12 anos	1
#ap 1. has 2. tabagismo cerca de 10anos em	0
+ esclerodactilia) # nega tabagismo em uso de: mtx	0
- # hábitos: nega tabagismo e etilismo # medicamentos	1
- dislipidemia - nega tabagismo # medicações em uso:	1
- dlp - nega tabagismo e etilismo # exames:	1
- dm 2 - tabagismo - carcinoma basocelular -	0
- oriento cessar o tabagismo e o etilismo; -	1
- oriento paciente cessar tabagismo (2mços/dia) - oriento manutenção	0
-oriento riscos associados ao tabagismo e gestação; -ofereço psico	0
03 meses -oriento cessar tabagismo	1
1 mês - cessar tabagismo	1
2004 (6 meses) 2) tabagismo atual => 1 ano-maço	0
70 anos # comorbidades: tabagismo (1 maço e meio	1
8 anos / nega tabagismo ou etilismo	1
8 anos / nega tabagismo ou etilismo # uso	1
91% (sem história de tabagismo prévio para pensarmos em	0
a contraste + nega tabagismo ou etilismo prévio #	1
a noite nega dm tabagismo balconista	0
abstinência durante internação! nega tabagismo desde 2003. fez tratamento	0
alergias nega etilismo ou tabagismo	1
algias no momento alergias, tabagismo etilismo, hipertensão diabetes refere	1
ambulatorio de cessação de tabagismo - agendo retornos: >	1
anos de historia de tabagismo (5 cigarros de corda	0
antecedentes pessoais e comorbidades: tabagismo (8 cigarros/dia), etilismo (final	1

5.1 Pré-processamento Textual

- Tokenização: Divisão dos textos em palavras ou tokens.
- Remoção de stopwords: Eliminação de palavras irrelevantes para a análise, como preposições e artigos.
- Normalização: Conversão dos textos para letras minúsculas e remoção de acentuações.
- Vetorização: Utilização de duas abordagens principais:
 - TF-IDF (Term Frequency-Inverse Document Frequency)
 - Embeddings de LLMs (Large Language Models)

5.2 Representação Vetorial

TF-IDF (Term Frequency-Inverse Document Frequency)

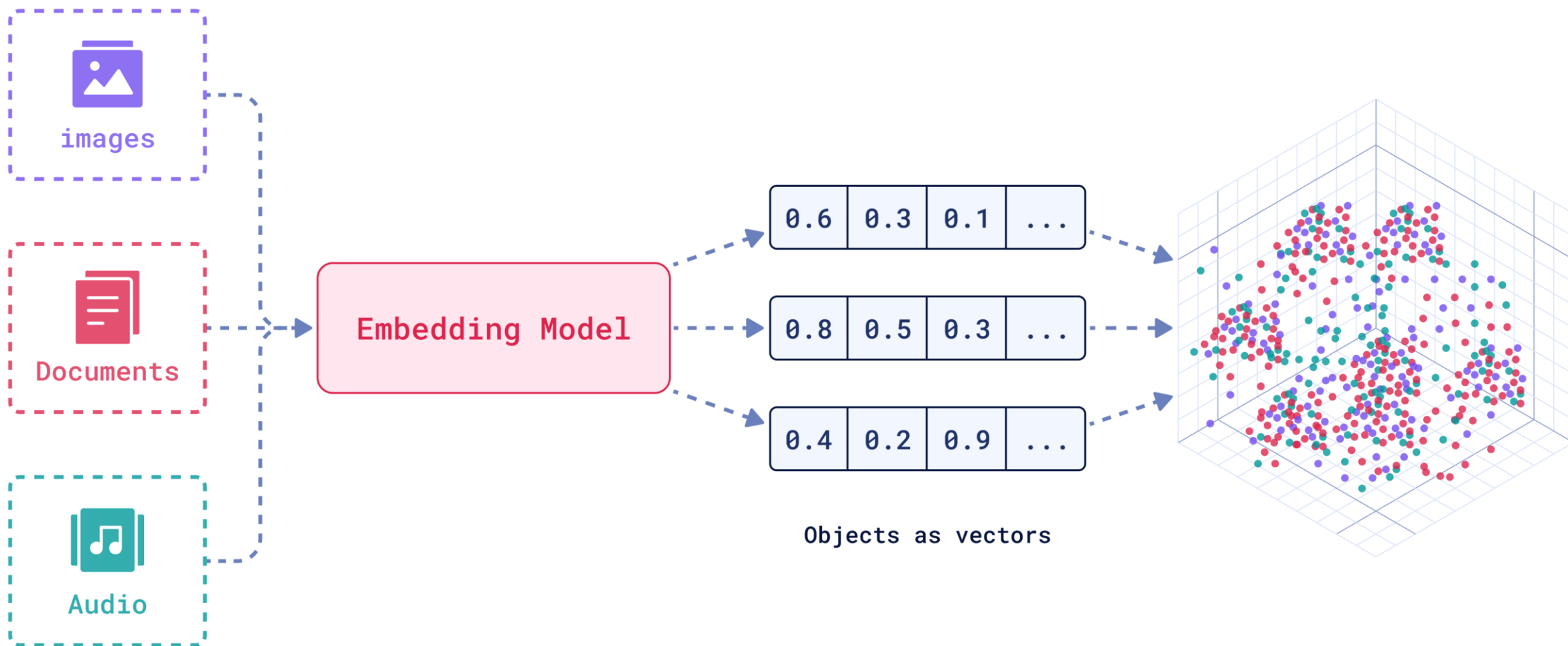
- Transformação dos textos em vetores esparsos que destacam a importância relativa de termos frequentes e raros no conjunto de dados.

Embeddings de LLMs (Large Language Models)


- Geração de vetores densos que capturam as relações semânticas profundas entre palavras e contextos.

Embeddings Vetoriais

- A qualidade das representações vetoriais impulsiona o desempenho
- São sobre semântica, "uma palavra é conhecida pela companhia que mantém"
- Depois que os vetores são armazenados, podemos usar suas propriedades espaciais para realizar pesquisas de vizinhos mais próximos
- Essas pesquisas recuperam itens semanticamente semelhantes com base em quão próximos eles estão neste espaço
- São usadas redes neurais para atribuir valores numéricos aos dados de entrada, de forma que dados semelhantes tenham valores semelhantes



5.3 Modelos de Classificação

- Regressão Logística
 - Árvore de Decisão
 - Implementação e Treinamento dos Modelos
 - Avaliação dos Modelos
- 

Resultados

Resultados

- Importância da Integração de Dados não Estruturados nas Predições Clínicas
- Processamento Inicial e Categorização de Tabagismo
- Impacto da Categorização no Machine Learning
- Aplicação de Modelos de Machine Learning

Resultados

1. Desafios na Integração de Dados Não Estruturados:

- Dados críticos sobre **tabagismo**, **etilismo** e **obesidade** estavam em campos de texto livre, dificultando a categorização.
- Notas clínicas apresentam vocabulário variado e ausência de estrutura, limitando sua utilização em modelos preditivos.

2. Importância de Dados Não Estruturados:

- Registros em texto livre capturam nuances importantes do perfil clínico e comportamentos dos pacientes.
- Integração com técnicas de **Processamento de Linguagem Natural (PNL)** potencializa a análise preditiva em saúde.
- Transformação desses dados em variáveis estruturadas melhora a compreensão de fatores de risco.

Resultados

3. Abordagem de Categorização Avançada:

- Utilização de LLMs e **técnicas de SQL avançado** para categorizar descrições complexas.
- Refinamento de expressões relacionadas ao tabagismo, como "cessou tabagismo" e "ex-tabagista".
- Criação de colunas categóricas para análise detalhada no BigQuery, incluindo **análise temporal** e de **contexto**.

4. Impacto no Machine Learning:

- Dados categorizados transformados em um target binário ('smoke' e 'nosmoke').
- Melhorou a precisão das predições de MACE, otimizando o uso de dados clínicos textuais.
- Demonstração clara do impacto do pré-processamento na eficácia dos modelos.

Aplicação de Modelos de Machine Learning

Regressão Logística

- Simplicidade e capacidade de resolver problemas de **classificação binária**
- Foi treinada com **vetores TF-IDF** e **embeddings de LLMs**, alcançando uma acurácia de 91,6%.
- Foi eficaz em **associar palavras-chave relacionadas ao tabagismo** à classificação final de fumante ou não fumante
- Adequado para futuros estudos multicêntricos, onde a padronização e a simplicidade na interpretação são essenciais

Árvores de Decisão

- Usado visando **capturar interações mais complexas nos textos**
- Permite **segmentar informações textuais de maneira não linear**
- Acurácia (89,4%) ligeiramente inferior à da regressão logística, porém a capacidade interpretativa do modelo e sua flexibilidade ao lidar com descrições indiretas ou ambíguas tornam-no valioso para a análise clínica.

Avaliação de Desempenho

Métricas utilizadas:

- Precisão, Recall e F1-score: Avaliam o equilíbrio entre classificações corretas e erros.
- Matriz de Confusão: Destaca a distribuição de predições corretas e incorretas (Tabelas 1 e 2).
- Curva ROC: Mede a capacidade discriminativa dos modelos.

Resultados Destacados

- Representação TF-IDF: Mais eficaz para textos padronizados e curtos.
- Resultados consistentes na matriz de confusão com alta precisão (92%) e acurácia (91,6%).

Representação com Embeddings LLM:

- Melhor em textos complexos, capturando nuances semânticas.
- AUC da curva ROC foi moderada (0.68 no treino) e próxima de aleatória no teste (0.47), indicando sobreajuste.

Avaliação dos Modelos e Métricas de Desempenho

Tabela 1

Matriz de Confusão da Regressão Logística	
Previsão: Fumante	Previsão: Não Fumante
Fumante: 94	6
Não Fumante: 5	95

Ambos os modelos tiveram uma alta taxa de sensibilidade (recall), com baixas taxas de falsos negativos, o que é crucial no contexto clínico para evitar a subestimação do risco em pacientes fumantes.

Tabela 2

Desempenho dos Modelos			
Modelo	Acurácia	Precisão	Recall
Regressão Logística	91,6%	92%	90%
Árvore de Decisão	89,4%	90%	88%

Acima temos o desempenho do modelo em termos de classificação correta de exemplos individuais.

Abordagem Híbrida: TF-IDF e Embeddings LLM

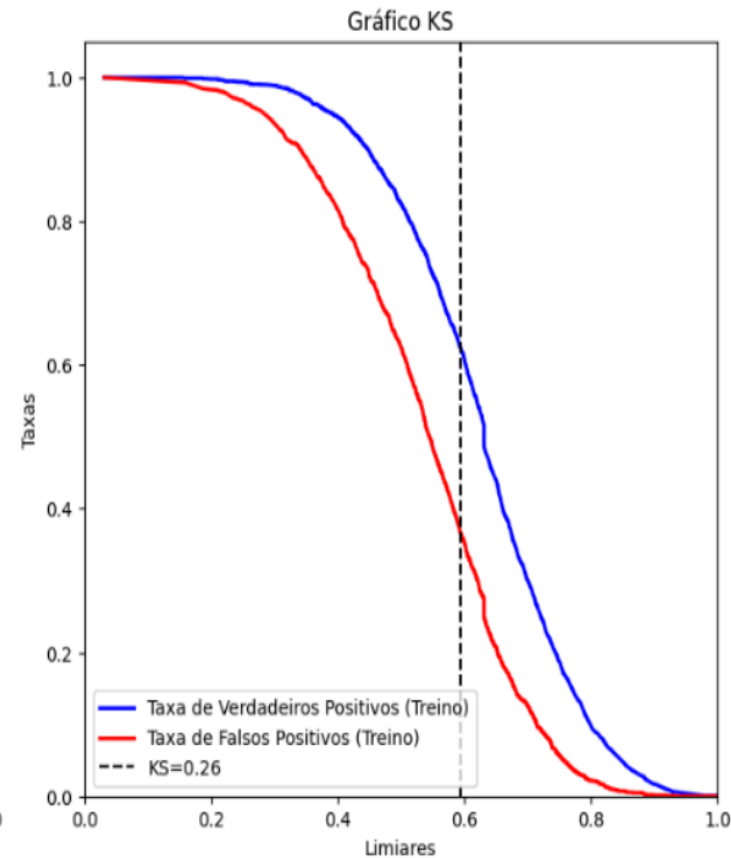
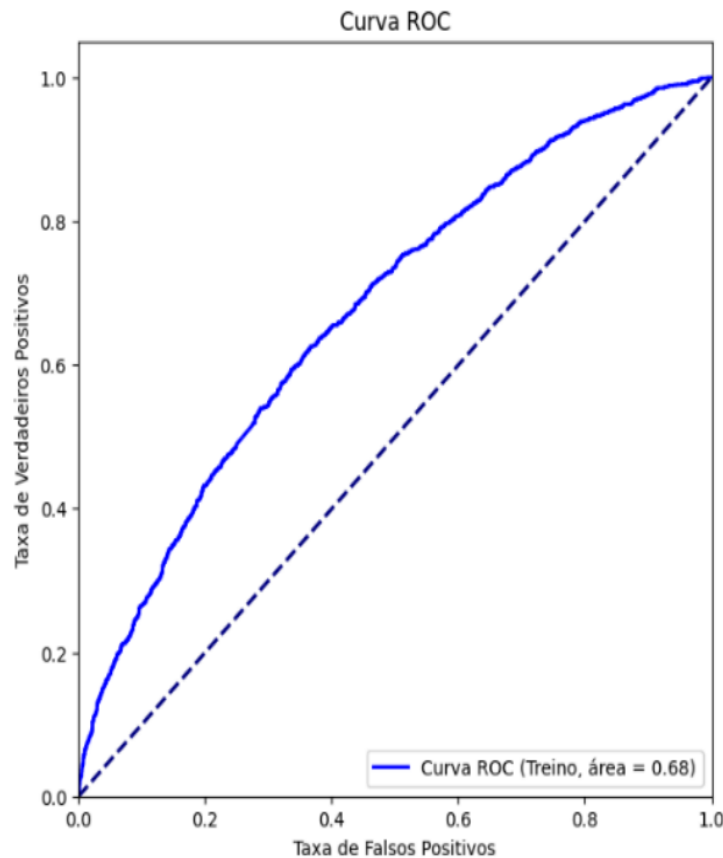
- Uma contribuição central deste estudo foi a contribuição de duas abordagens clássicas, como o TF-IDF, e métodos mais modernos, baseados em embeddings de LLMs
1. O texto na coluna CTU_INFORMACAO é pré-processado de forma mínima, priorizando a **coerência semântica**, sem aplicar transformações excessivas
 2. Foi utilizado o modelo paraphrase-multilingual-MiniLM-L12-v2 da biblioteca SentenceTransformer (eficaz em várias línguas) para gerar embeddings numéricos de alta dimensionalidade (384 dimensões) para cada frase ou sentença
 3. Estes embeddings são extraídos a partir do texto original para preservar a riqueza semântica, fornecendo ao modelo representações numéricas
 4. Diferente dos embeddings, o TF-IDF gera representações numéricas focadas na frequência e importância dos termos

Análise Comparativa: TF-IDF e Embeddings LLM

Dimensão 0	Dimensão 1	...	Dimensão 382	Dimensão 383	Classe (Y)
-0.064247	0.178107	...	-0.069604	0.214405	1
0.108418	0.470138	...	-0.236791	0.328719	1
0.051831	0.381500	...	-0.011956	0.135567	0
-0.025718	0.445989	...	0.272254	0.277455	0
0.141891	0.311944	...	0.197493	0.154731	1

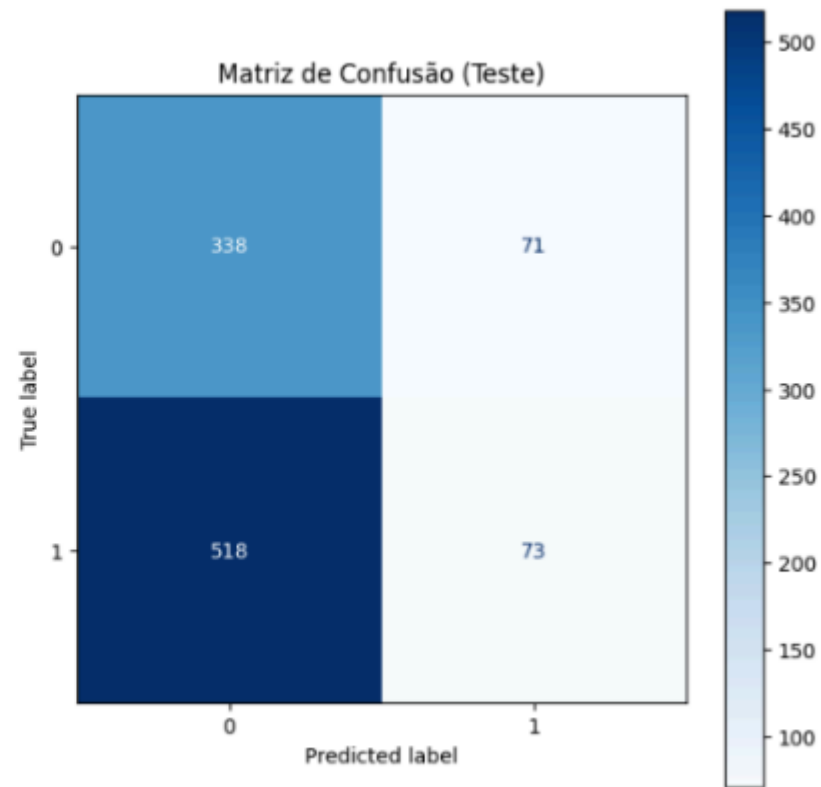
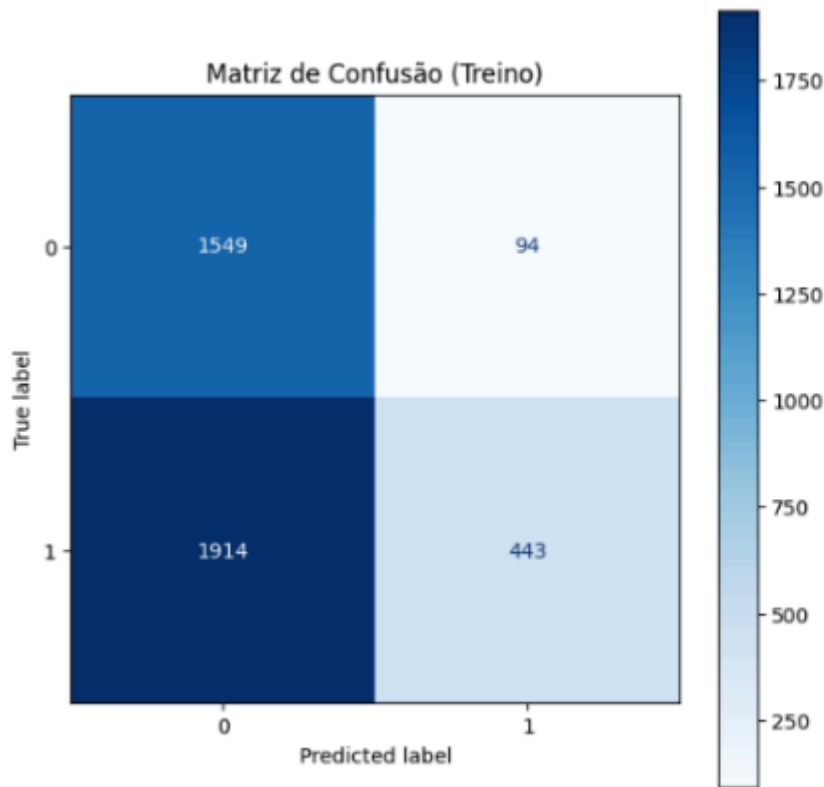
- Cada coluna numerada de 0 a 383 corresponde aos componentes dos embeddings gerados, enquanto a última coluna (Y) indica a classe-alvo associada ao problema de classificação, sendo binária (0 ou 1)
- Esses embeddings serão então utilizados posteriormente como entrada em modelos de aprendizado de máquina, como no caso da predição de eventos cardiovasculares adversos maiores (MACE) no PRECARE-ML

Estudo dos modelos gerados



- Treinamento: AUC de 0.68 (moderada capacidade de discriminação).
- Teste: AUC caiu para 0.47 (próximo ao aleatório).
- Problema: Indício de sobreajuste no modelo.
- Causa: Aprendizagem específica do treino, sem boa generalização

Estudo dos modelos gerados



- Métrica usada: Matriz de confusão
- Acertos: 338 instâncias corretas da classe negativa
- Erro crítico: 518 falsos negativos (MACE)
- Problema: Dificuldade em identificar casos positivos
- Impacto: Risco de falha em medidas preventivas

Discussão

- A utilização dessa abordagem híbrida—ao combinar embeddings para capturar a semântica e TF-IDF para capturar a importância e frequência das palavras ofereceu um modelo robusto, capaz de explorar diferentes dimensões dos dados textuais sobre tabagismo
- Esse estudo explorou tanto a riqueza semântica quanto às frequências textuais das informações encontradas em campos de texto livre de prontuários eletrônicos

Conclusão

- **Avanço:** Dados não estruturados, como tabagismo, foram transformados em variáveis estruturadas, utilizáveis em modelos preditivos.
- **Desempenho:** Apesar de métricas moderadas, a abordagem mostrou-se promissora devido à sua escalabilidade e potencial de evolução.
- **Técnicas aplicadas:** A combinação de embeddings e TF-IDF capturou tanto a semântica quanto a relevância nos textos médicos.
- **Automatização:** A estruturação de textos livres possibilitou análises personalizadas e intervenções clínicas preventivas mais eficazes.
- **Uso de LLMs:** Modelos LLM tokenizaram textos e geraram vetores numéricos, aprimorando a precisão na categorização de tabagismo.
- **Potencial:** A metodologia é aplicável a outros fatores de risco, como obesidade e etilismo, ampliando o alcance das análises.
- **Futuro:** Explorar novas variáveis, refinar técnicas e aprimorar modelos para aumentar a precisão e personalização dos cuidados de saúde.

Referências

1. Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., & Nichols, M. (2016). Cardiovascular disease in Europe: Epidemiological update 2016. In *European Heart Journal* (Vol. 37, Issue 42). <https://doi.org/10.1093/eurheartj/ehw334>.
2. Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., Cooney, M. T., Corrà, U., Cosyns, B., Deaton, C., Graham, I., Hall, M. S., Hobbs, F. D. R., Løchen, M. L., Löllgen, H., Marques-Vidal, P., Perk, J., Prescott, E., Redon, J., ... Gale, C. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice. In *European Heart Journal* (Vol. 37, Issue 29). <https://doi.org/10.1093/eurheartj/ehw106>.
3. Zhang X, Wang L, Miao S, Xu H, Yin Y, Zhu Y, et al. Analysis of treatment pathways for three chronic diseases using OMOP CDM. *J Med Syst*. 2018;42(12).
4. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP a scoping review. In: *Studies in Health Technology and Informatics*. IOS Press BV; 2021. p. 95-103.
5. OHDSI. 2020_data_network [Internet]. Observational Healthcare Data Sciences and informatics (OHDSI). 2021 [cited 2022 Nov 10].https://www.ohdsi.org/web/wiki/doku.php?id=resources:2020_data_network
6. Steyerberg, E. et al. 2010. Assessing the Performance of Prediction Models: A Framework for Some Traditional and Novel Measures. *Epidemiology*, 21(1).
7. Moons, K. G. et al. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 6;162(1).
8. Vickers, A., and Elkin, E. 2008. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*, 26(6).