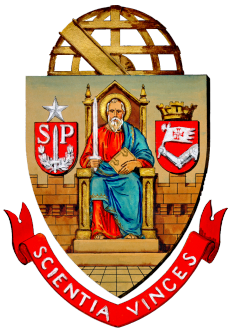

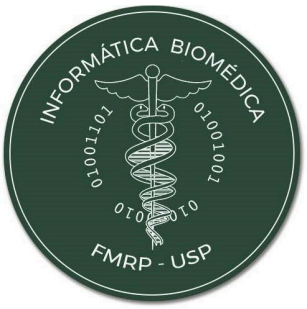

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

CURSO DE INFORMÁTICA BIOMÉDICA

MAYARA MARTINS PERRONI

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

RIBEIRÃO PRETO - SP

2024

MAYARA MARTINS PERRONI

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

Trabalho de Conclusão de Curso apresentado para obtenção de título de Bacharel em Informática Biomédica pela Faculdade de Medicina de Ribeirão Preto/USP

Área de Concentração: Informática Biomédica

Orientador: Prof. Dr. Paulo Mazzoncini de Azevedo Marques

RIBEIRÃO PRETO - SP

2024

RESUMO

O avanço da tecnologia da informação em saúde possibilitou a coleta e armazenamento de grandes volumes de dados em formatos variados, essenciais para apoiar a tomada de decisões clínicas. No entanto, a heterogeneidade desses dados pode comprometer o desempenho de modelos de aprendizado de máquina (AM), causando sobreajuste (overfitting) e limitando a generalização dos resultados. Para lidar com esse desafio, foi implementado no BigQuery um processo de Extraction, Transformation and Loading (ETL), que padronizou dados provenientes do Hospital das Clínicas (HC) de registros eletrônicos de saúde (RES), utilizando o Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM), desenvolvido pela comunidade Observational Health Data Sciences and Informatics (OHDSI). Esse processo garantiu a uniformidade dos dados, permitindo sua integração em diferentes instituições de saúde. O objetivo foi desenvolver e testar um classificador para identificar fumantes a partir de textos livres extraídos de prontuários eletrônicos. Para isso, foram aplicadas duas abordagens de aprendizado de máquina: Large Language Models (LLM) e Term Frequency-Inverse Document Frequency (TF-IDF). Os textos, padronizados no OMOP-CDM, passaram por etapas de pré-processamento, categorização e análise comparativa. O LLM focou na interpretação semântica e contextual dos textos, enquanto o TF-IDF identificou padrões de palavras-chave associados ao tabagismo. Após os testes, ambos os modelos foram avaliados em termos de precisão, recall e capacidade de generalização, considerando a variabilidade dos dados. Os resultados indicam que o LLM oferece uma abordagem mais eficaz para categorizar fatores de risco em dados não estruturados, proporcionando uma análise mais profunda dos textos livres. Por outro lado, o TF-IDF se mostrou vantajoso em contextos com padrões de texto mais controlados e repetitivos. Este estudo contribui para o desenvolvimento de classificadores automáticos que podem ser integrados a sistemas de prontuários eletrônicos, aprimorando a categorização de informações críticas, como o histórico de tabagismo, e servindo de base para futuras análises de fatores de risco em registros padronizados.

Palavras-Chave: Processamento de Linguagem Natural, OMOP-CDM, Aprendizado de Máquina, Classificação de Fumantes, TF-IDF, LLM, Registros Eletrônicos de Saúde, Tabagismo

ABSTRACT

The advancement of health information technology has enabled the collection and storage of large volumes of data in various formats, essential for supporting clinical decision-making. However, the heterogeneity of this data can compromise the performance of machine learning (ML) models, causing overfitting and limiting the generalization of results. To address this challenge, an Extraction, Transformation, and Loading (ETL) process was implemented in BigQuery, standardizing data from electronic health records (EHR) of the Hospital das Clínicas (HC), using the Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM), developed by the Observational Health Data Sciences and Informatics (OHDSI) community. This process ensured data uniformity, enabling integration with data from other health institutions. The objective was to develop and test a classifier to identify smokers from free-text entries extracted from electronic medical records. For this purpose, two machine learning approaches were applied: Large Language Models (LLM) and Term Frequency-Inverse Document Frequency (TF-IDF). The texts, standardized in the OMOP-CDM format, underwent preprocessing, categorization, and comparative analysis. The LLM focused on the semantic and contextual interpretation of the texts, while the TF-IDF highlighted keyword patterns associated with smoking. The models were evaluated in terms of precision, recall, and generalization capacity, considering the variability of the data. The results indicated that the LLM provides a more effective approach for categorizing risk factors in unstructured data, offering deeper analysis of free-text entries. On the other hand, the TF-IDF proved advantageous in contexts with more controlled and repetitive text patterns. This study contributes to the development of automatic classifiers that can be integrated into electronic medical record systems, enhancing the categorization of critical information, such as smoking history, and serving as a foundation for future analyses of risk factors in standardized records.

Keywords: Natural Language Processing, OMOP-CDM, Machine Learning, Smoking Classification, TF-IDF, LLM, Electronic Health Records, Smoking

Lista de Figuras

Lista de Tabelas

SUMÁRIO

1. INTRODUÇÃO

Os eventos cardiovasculares adversos maiores (Major Adverse Cardiac Events - MACE) têm um impacto significativo na saúde pública em todo o mundo. MACE, que incluem infarto do miocárdio, acidente vascular cerebral e morte cardiovascular, são uma das principais causas de mortalidade globalmente, representando uma grande preocupação para sistemas de saúde e comunidades científicas. Diversos fatores de risco, como hipertensão, diabetes, dislipidemia e tabagismo, estão associados à maior probabilidade de ocorrência desses eventos. No entanto, a predição precoce desses eventos continua sendo um desafio, especialmente em ambientes clínicos complexos, onde uma combinação de fatores de risco e condições médicas coexistem. A inteligência artificial (IA) é considerada uma tecnologia de fronteira revolucionária e tem se tornado um interesse global de pesquisa na área da medicina, incluindo a medicina cardiovascular (Shu, Ren & Song, 2021). O uso de modelos de aprendizado de máquina (AM) tem se mostrado promissor na predição de MACE, integrando dados estruturados e não estruturados de prontuários eletrônicos de saúde (EHR) e oferecendo novas oportunidades para aprimorar o cuidado preventivo (Polat Erdeniz et al., 2023).

Nesse contexto, surge o projeto multicêntrico PRE-CARE ML (Prevendo eventos cardiovasculares usando aprendizado de máquina), financiado pela FAPESP (#2021/06137-4), que é uma colaboração internacional entre a Medical University of Graz e Styrian Hospital Association (Áustria), o Hasso Plattner Institut - University of Potsdam (Alemanha), o Karolinska Institutet (Suécia) e a Faculdade de Medicina de Ribeirão Preto (FMRP), em conjunto com o Hospital das Clínicas de Ribeirão Preto (Brasil). O projeto utiliza dados de pacientes hospitalizados por condições não cardiológicas para prever o risco de desenvolvimento de MACE a partir da extração de features dos EHR. O objetivo é fornecer, no momento da alta, um escore preditivo que apoie o paciente e seus médicos na identificação precoce de riscos, possibilitando intervenções preventivas direcionadas. Essa abordagem amplia a aplicabilidade de modelos preditivos ao englobar pacientes fora do espectro típico de atenção cardiovascular direta, promovendo um cuidado mais holístico e personalizado.

Apesar de já existir uma extensa base de dados sobre fatores de risco como hipertensão, diabetes e tabagismo, a predição precoce desses eventos em ambientes clínicos continua sendo um desafio. Isso se deve à complexidade de

integrar fatores estruturados e, principalmente, dados não estruturados, como descrições livres em registros eletrônicos de saúde (EHR), que muitas vezes contêm informações essenciais para a previsão de MACE, mas que permanecem subutilizadas.

Técnicas de aprendizado de máquina oferecem uma abordagem promissora, especialmente quando aplicadas à categorização e vetorização de campos de texto livre dos prontuários. Essas técnicas permitem que informações anteriormente não estruturadas, como sintomas descritos por médicos e fatores de risco mencionados de forma textual, sejam organizadas e transformadas em variáveis estruturadas e utilizáveis para modelos preditivos. No âmbito do projeto essas técnicas são aplicadas para integrar dados estruturados e não estruturados em um modelo preditivo robusto. A proposta é fornecer para o paciente e para seu médico de referência, ou médico de família, um escore preditivo de risco de ocorrência de algum desfecho de MACE no prazo de até cinco anos. Considerando que esses eventos estão fortemente relacionados com o estilo de vida, a intenção é fornecer um alerta visando possibilitar intervenções preventivas voltadas para a minimização do risco.

O objetivo central do trabalho aqui apresentado, realizado no contexto desse projeto multicêntrico, foi a estruturação de texto livre contendo informações sobre tabagismo presente nos registros eletrônicos de saúde extraídos dos sistemas de informação do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. A vetorização dos dados categorizados, passo-chave para integrar informações não estruturadas ao modelo OMOP-CDM, facilita uma análise preditiva mais robusta e representativa. O processo inclui a tokenização de texto, utilizando abordagens de processamento de linguagem natural (PLN), como TF-IDF (Term Frequency-Inverse Document Frequency), e a aplicação de técnicas mais modernas, como Modelos de Linguagem de Grande Escala (Large Language Models - LLMs), para refinar ainda mais a categorização e otimização dos dados.

Essa metodologia busca não apenas melhorar a precisão das previsões de MACE, mas também proporcionar uma ferramenta automatizada para o tratamento contínuo de texto livre em dados clínicos, garantindo uma integração mais eficaz dessas informações nos sistemas de saúde.

1.1. Categorização de Fatores de Risco e Dados em Texto Livre

Um dos maiores desafios na predição de eventos MACE a partir de dados

extraídos de EHR é o tratamento de dados não estruturados, como descrições de fatores de risco e sintomas relatados em campos de texto livre. Informações como tabagismo, etilismo, obesidade e sintomas específicos são geralmente registradas de forma textual, sem um formato padronizado. Isso complica a análise e integração dessas informações com outros dados estruturados, como resultados laboratoriais e históricos médicos.

Na construção de modelos preditivos em saúde, como aqueles voltados para a predição de eventos cardiovasculares adversos, a inclusão de dados textuais não estruturados como uma feature (ou variável) importante tem sido uma área de crescente interesse. 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.

A principal dificuldade em utilizar campos de texto livre em modelos de aprendizado de máquina é a sua natureza não padronizada. Termos que descrevem fatores de risco, como o consumo de álcool ou tabagismo, podem ser expressos de várias formas pelos profissionais de saúde. Isso exige o uso de técnicas avançadas de Processamento de Linguagem Natural (PLN) para transformar esses textos em representações que possam ser interpretadas por modelos preditivos.

Ao incluir essas features derivadas de texto livre, o modelo se torna capaz de utilizar informações não estruturadas que, de outra forma, seriam negligenciadas em uma abordagem exclusivamente baseada em dados estruturados. Isso pode aumentar a acurácia preditiva, uma vez que esses textos frequentemente contêm detalhes ricos que complementam os dados tabulares, como resultados laboratoriais e históricos de tratamentos. A incorporação de campos de texto livre como features, portanto, representa uma forma de enriquecer os modelos preditivos de MACE, potencializando sua capacidade de identificar pacientes em risco com maior precisão. Isso se alinha com o objetivo mais amplo de personalizar intervenções clínicas com base em um entendimento mais completo do estado de saúde do paciente, capturado tanto em dados estruturados quanto não estruturados.

1.2. Otimização da Categorização e Vetorização

A categorização e vetorização de dados textuais representam passos fundamentais para a transformação de informações não estruturadas em formatos

adequados para modelos de aprendizado de máquina. No contexto de predição de eventos clínicos, como os MACE, essa otimização visa garantir que dados provenientes de prontuários médicos, geralmente registrados de forma textual, possam ser utilizados de maneira eficiente.

A categorização é o processo pelo qual dados textuais são classificados em grupos ou categorias que representam fatores de risco ou sintomas relevantes. Isso pode incluir, por exemplo, a classificação de pacientes com base no seu histórico de tabagismo, etilismo, ou condições como obesidade. 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.

No entanto, 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. Técnicas como o TF-IDF (Term Frequency-Inverse Document Frequency) são amplamente utilizadas para essa tarefa, pois 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, evitando tanto a perda de informações quanto a introdução de ruídos. Um modelo de predição de MACE, por exemplo, pode se beneficiar de categorias bem definidas para fatores de risco, enquanto a vetorização precisa garantir que esses fatores sejam adequadamente ponderados, sem sobrecarregar o modelo com informações redundantes. No mais, é comum a aplicação de técnicas como redução dimensional para diminuir a complexidade dos vetores, assegurando que o modelo mantenha seu desempenho ao lidar com grandes volumes de dados. A regularização também desempenha um papel importante, evitando o overfitting ao modelo e assegurando que ele generalize bem para diferentes populações clínicas.

1.3. Aprendizado de Máquina na Saúde

O uso de modelos de aprendizado de máquina no campo da saúde tem crescido exponencialmente, com sua capacidade de analisar grandes volumes de dados e detectar padrões complexos. Em relação à predição de MACE, esses modelos permitem que diferentes fontes de dados, como históricos médicos,

resultados laboratoriais e dados demográficos, sejam integradas para gerar predições mais precisas. 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 (Polat Erdeniz et al., 2023). A relevância científica desta abordagem reside em sua capacidade de superar as limitações dos métodos tradicionais de avaliação de risco, como o escore de Framingham, que não consegue captar plenamente a complexidade das interações entre múltiplos fatores de risco.

Redes neurais artificiais são compostas por camadas de neurônios artificiais que processam informações de maneira semelhante ao cérebro humano. Elas são particularmente eficazes em identificar padrões complexos em grandes conjuntos de dados. As máquinas de vetores de suporte (SVM), por outro lado, são algoritmos de aprendizado supervisionado que encontram o hiperplano ótimo que separa diferentes classes de dados. Ambas as técnicas têm se mostrado eficazes na predição de MACE devido à sua capacidade de lidar com dados de alta dimensionalidade e complexidade.

O OMOP-CDM tem se estabelecido como um modelo padrão para a integração de dados clínicos de diferentes instituições, preparando o terreno para estudos observacionais multicêntricos. A padronização dos dados em um formato comum é essencial para assegurar que os modelos de aprendizado de máquina possam ser aplicados de maneira consistente em diversos contextos e populações..

A padronização de dados implica a transformação de informações heterogêneas em um formato uniforme, o que apresenta desafios devido às discrepâncias nos sistemas de codificação, terminologias médicas e práticas de registro de dados entre diferentes instituições. No entanto, a adoção do OMOP-CDM facilita essa padronização, criando um ambiente propício para a comparação e integração de dados provenientes de múltiplos centros de saúde.

Embora a maioria dos modelos preditivos de MACE se baseie em dados estruturados, como exames laboratoriais e diagnósticos codificados, há uma quantidade substancial de informações valiosas em dados não estruturados, como notas clínicas e relatórios médicos. Estes 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). O Processamento de Linguagem Natural (PLN) é uma subárea da inteligência artificial que se concentra na interação entre computadores e a linguagem humana. Técnicas de PLN

permitem que os sistemas compreendam, interpretem e gerem linguagem humana de maneira significativa. No contexto da saúde, o PLN é utilizado para extrair informações valiosas de dados não estruturados, como notas de prontuários eletrônicos, relatórios médicos e comunicações entre profissionais de saúde. Técnicas de PLN, como a extração de entidades nomeadas (NER), análise de sentimentos e reconhecimento de padrões, são usadas para entender o contexto e a semântica de textos médicos, permitindo uma análise mais profunda e precisa.

"O processamento de linguagem natural (PNL) demonstrou repetidamente sua viabilidade para desbloquear evidências enterradas em narrativas clínicas" (SPASIC; NENADIC, 2020). Isso é reafirmado pelo fato que este método pode ajudar a identificar informações críticas que podem ser negligenciadas em análises tradicionais de dados estruturados. Por exemplo, sintomas descritos em notas clínicas podem fornecer pistas adicionais para diagnósticos mais precisos. A capacidade de integrar dados estruturados e não estruturados permite uma visão mais holística do estado de saúde do paciente. Isso é crucial para a predição de MACE, onde múltiplos fatores de risco e condições coexistem. Sistemas de suporte à decisão clínica (CDSS) que utilizam PLN podem fornecer recomendações baseadas em uma análise abrangente dos dados do paciente, incluindo notas clínicas e históricos médicos. Isso pode melhorar a tomada de decisões por parte dos profissionais de saúde.

O OMOP-CDM tem se estabelecido como um modelo padrão para estudos observacionais multicêntricos, permitindo a integração de dados clínicos de diferentes instituições de maneira harmonizada. A padronização dos dados em um formato comum é vital para garantir que os modelos de aprendizado de máquina (AM) possam ser aplicados de forma consistente em diferentes contextos e populações. No contexto deste estudo, dados de três centros médicos foram convertidos para o modelo OMOP, abrangendo informações detalhadas sobre diagnósticos, procedimentos, medicamentos e outros fatores relacionados à saúde cardiovascular (Polat Erdeniz et al., 2023).

A padronização de dados envolve a transformação de dados heterogêneos em um formato uniforme, o que pode ser desafiador devido às diferenças nos sistemas de codificação, terminologias médicas e práticas de registro de dados entre diferentes instituições. No entanto, a utilização do OMOP-CDM facilita essa padronização, permitindo a comparação e a integração de dados de múltiplos centros de saúde. Não obstante, futuras direções do trabalho podem incluir a aplicação de técnicas avançadas de vetorização, como Word Embeddings (por

exemplo, Word2Vec e BERT), que podem melhorar a captura da semântica nas descrições textuais. A adoção dessas técnicas permitirá uma análise mais profunda dos dados clínicos, promovendo uma compreensão mais rica das informações e, assim, contribuindo para a eficácia dos modelos de AM em um contexto multicêntrico.

Nesse contexto, o presente trabalho abordou a padronização de dados como uma etapa fundamental para mitigar essas limitações. Embora a padronização tenha sido realizada com base em um único centro de dados, ela estabelece uma base sólida para futuras aplicações multicêntricas. Assim, o uso de técnicas avançadas de vetorização, como Word Embeddings, visa capturar nuances semânticas que podem enriquecer a interpretação dos dados clínicos e melhorar o desempenho dos modelos de ML. Com isso, a aplicação dessas técnicas, associada à padronização dos dados, não apenas pode otimizar o desempenho dos modelos, mas também oferece uma base para que, em trabalhos futuros, seja possível validar e generalizar os modelos em contextos multicêntricos.

Em suma, embora os desafios da variabilidade dos dados clínicos entre diferentes centros persistam, nossa abordagem busca contribuir para a construção de modelos de ML mais robustos e generalizáveis. Ao integrar técnicas padronizadas e inovadoras, esperamos que este trabalho ajude a pavimentar o caminho para previsões mais precisas e personalizadas, proporcionando melhores cuidados de saúde em larga escala.

2. PROBLEMA

As 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. A aterosclerose, condição central nas DCVs, é responsável por infarto do miocárdio, doença cardíaca isquêmica e acidente vascular cerebral, comprometendo gravemente a qualidade de vida dos pacientes. Para minimizar esses impactos, é essencial identificar precocemente indivíduos com alto risco de eventos cardiovasculares adversos maiores (MACE), o que possibilita a implementação de intervenções preventivas mais eficazes.

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. Embora úteis, essas abordagens carecem de precisão e personalização. O aprendizado de máquina (ML) surge como uma alternativa promissora, capaz de integrar uma gama mais ampla de variáveis e detectar padrões complexos. Modelos de ML, como redes neurais e árvores de decisão, têm demonstrado potencial para superar os métodos convencionais na predição de MACE, mas a validação desses modelos em contextos clínicos reais ainda enfrenta obstáculos, especialmente pela variabilidade dos dados clínicos entre diferentes centros e populações.

Este estudo, inserido no projeto PRE-CARE ML, propõe investigar uma possível abordagem para superar esses desafios, explorando ao máximo dados não estruturados presentes nos registros eletrônicos de saúde (EHRs), especialmente aqueles que tratam de hábitos de vida como tabagismo, etilismo e obesidade. Muitas dessas informações, registradas em texto livre, são tradicionalmente descartadas pela dificuldade de extração e categorização. Aplicando modelos de linguagem de grande escala (LLMs), o estudo converte esses dados em categorias estruturadas, facilitando uma integração fluida com variáveis estruturadas nos modelos preditivos de ML. Essa metodologia aprimora significativamente a abrangência e a precisão dos modelos de predição de MACE, permitindo que indicadores importantes, antes subutilizados, contribuam para uma análise mais precisa e personalizada do risco cardiovascular.

3. HIPÓTESE

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.

4. OBJETIVO

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.

5. MATERIAL E MÉTODOS

O estudo desenvolvido foi submetido e aprovado pelo Comitê de Ética em Pesquisa (CEP) da Plataforma Brasil, garantindo o cumprimento dos princípios éticos. A pesquisa visa a predição de MACE (Eventos Cardiovasculares Adversos Maiores) em pacientes utilizando modelos de aprendizado de máquina e redes neurais, com dados padronizados no OMOP-CDM (Observational Medical Outcomes Partnership – Common Data Model). A coleta e manipulação de dados foram conduzidas de acordo com as diretrizes de segurança e anonimização dos dados de saúde, provenientes de três centros multicêntricos.

5.1 . Materiais Utilizados

Foram utilizados dados eletrônicos de saúde (EHR) de três multicentros, anonimizados e padronizados no OMOP-CDM. As coortes foram organizadas de acordo com a padronização do OHDSI (Observational Health Data Sciences and Informatics). Os pacientes incluídos no estudo eram adultos diagnosticados com doenças cardiovasculares, com histórico clínico disponível nas bases dos multicentros.

Além dos dados clínicos estruturados, foram utilizados campos livres referentes a hábitos de vida (etilismo, tabagismo, obesidade), que foram processados e categorizados por meio de LLMs (Large Language Models) para permitir a integração com os demais fatores estruturados.

5.1.1 . *Softwares* utilizados

A infraestrutura tecnológica adotada para este projeto foi crucial para a coleta, processamento e análise de dados. As ferramentas e plataformas utilizadas incluíram:

BigQuery (Google Cloud Platform): Foi a solução escolhida para armazenamento e processamento dos dados anonimizados, organizados no formato OMOP-CDM. A utilização do BigQuery permitiu consultas rápidas e eficientes sobre grandes volumes de dados, facilitando a implementação de análises complexas.

Modelos de Aprendizado de Máquina em Python: Para a predição de MACE (Eventos Cardiovasculares Adversos Maiores), foram implementados modelos de aprendizado de máquina utilizando bibliotecas como TensorFlow e PyTorch. Esses

frameworks forneceram as ferramentas necessárias para a construção e treinamento de modelos complexos, permitindo uma análise aprofundada dos dados.

OHDSI ATLAS: Essa plataforma foi utilizada para a análise estatística e definição de coortes, facilitando a padronização dos dados em conformidade com o modelo OMOP. O OHDSI ATLAS permitiu realizar análises descritivas e explorar dados em um ambiente colaborativo.

Equipamentos: Em termos de equipamentos, a infraestrutura computacional foi predominantemente baseada em servidores em nuvem. A Google Cloud forneceu os recursos necessários para o processamento massivo dos dados e para o treinamento dos modelos de aprendizado de máquina. Essa abordagem em nuvem possibilitou flexibilidade, escalabilidade e eficiência, sem a necessidade de investimentos em hardware específico.

5.2. O Revisão Bibliográfica e Estudo de Procedimentos ETL

Inicialmente, foi realizada uma revisão bibliográfica detalhada sobre os processos de ETL (Extract, Transform, Load) e o modelo OMOP-CDM (Observational Medical Outcomes Partnership - Common Data Model). Esta revisão incluiu a análise da importância da padronização de dados para estudos multicêntricos, evidenciando como a harmonização e a estruturação de dados podem melhorar a integração e a análise de informações provenientes de diferentes fontes.

A revisão focou também em procedimentos de ETL utilizados em pesquisas multicêntricas que empregaram o OMOP-CDM. Esse estudo inicial permitiu identificar as melhores práticas e desafios associados à padronização de dados para a realização de análises comparativas entre diferentes centros de saúde. A tecnologia utilizada no projeto foi aplicada em etapas desenvolvidas durante a pesquisa, tal como apresentadas na Figura 1.

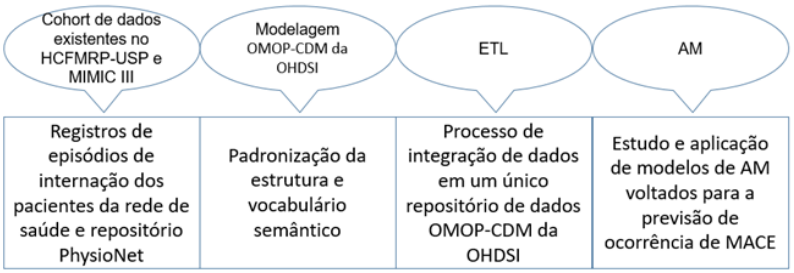

Figura 1: etapa dos materiais e métodos utilizados.

5.3. Implementação de ETL e Criação do Banco de Dados

Este projeto obteve aprovação do Comitê de Ética em Pesquisa do HCFMRP sob parecer número 6.071.163, com dispensa do termo de consentimento livre e esclarecido. O HCFMRP-USP, por meio do departamento de qualidade de dados, disponibilizou um volume de 3.952,9 Gigabytes de dados clínicos em arquivos no formato .csv. Cada arquivo foi processado para organização dos dados em colunas tabulares, separadas por delimitadores específicos. Durante os primeiros seis meses, o trabalho incluiu o apoio à organização das coortes de dados dos hospitais vinculados ao HCFMRP e do repositório MIMIC-III, com a implementação de processos de ETL (Extract, Transform, Load). Esses processos resultaram na criação de um ambiente de dados estruturado e padronizado, conforme o modelo OMOP-CDM da OHDSI.

A padronização envolveu a modelagem da estrutura e do vocabulário semântico, mapeando registros de pacientes e internações para garantir a interoperabilidade entre diferentes sistemas e bases de dados. A Figura 2 apresenta a representação esquemática parcial do fluxo de ETLs, enquanto a Figura 3 detalha as etapas principais do fluxo ETL, mostrando como cada script SQL contribuiu para a organização e integração dos dados no repositório OMOP-CDM.

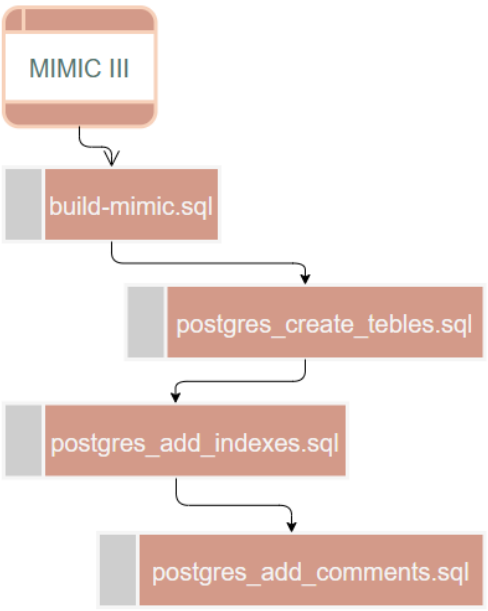

Figura 2: representação esquemática parcial do fluxo de ETLs.

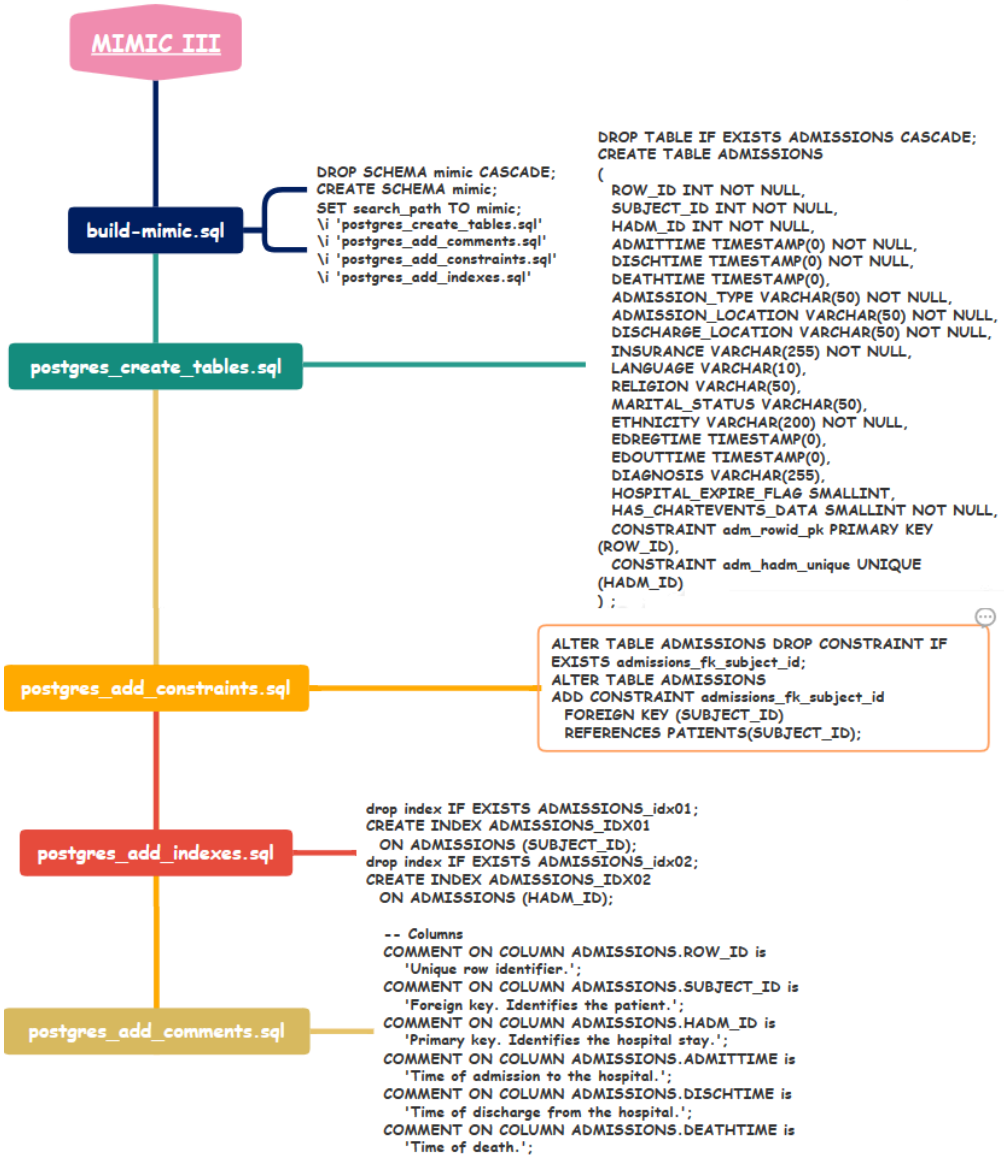

Figura 3: representação esquemática total do fluxo de ETLs.

5.4. Integração e Estruturação dos Dados

Após a organização inicial dos dados, o processo de integração foi realizado, consolidando as informações anonimizadas em um único repositório OMOP-CDM no Google BigQuery. Esse ambiente de banco de dados escalável possibilitou o armazenamento e a análise em larga escala de dados demográficos e clínicos dos pacientes, como idade, gênero e cidade de residência, fundamentais para os modelos de aprendizado de máquina (AM).

Com a estruturação dos dados no modelo OMOP-CDM, dois grupos distintos de pacientes foram definidos: o Grupo MACE, composto por pacientes que sofreram eventos cardiovasculares maiores, e o Grupo Controle, formado por

pacientes que não estavam sob tratamento cardiológico ou não apresentaram MACE durante o período avaliado. A inclusão desse grupo controle foi essencial para permitir comparações robustas e identificar padrões que diferenciam pacientes com alto risco de MACE daqueles sem evidências de tais eventos. Essa divisão possibilitou aos modelos de AM explorar tanto fatores de risco específicos quanto características que podem ser indicadores de um perfil de menor risco.

Foram também elaboradas tabelas complementares, incluindo registros de diagnósticos com classificação internacional de doenças (CID), que facilitaram a criação de coortes específicas para o estudo de MACE. A Figura 4 ilustra a estrutura parcial do banco de dados MIMIC-III no Postgres SQL, detalhando a tabela de admissões de pacientes internados, usada para investigação de eventos MACE. A Figura 5 apresenta o ambiente de desenvolvimento com integração entre Python e o banco de dados MIMIC-III no Microsoft Visual Studio Code (VSCode), que possibilitou a manipulação dos dados de forma organizada e eficiente.

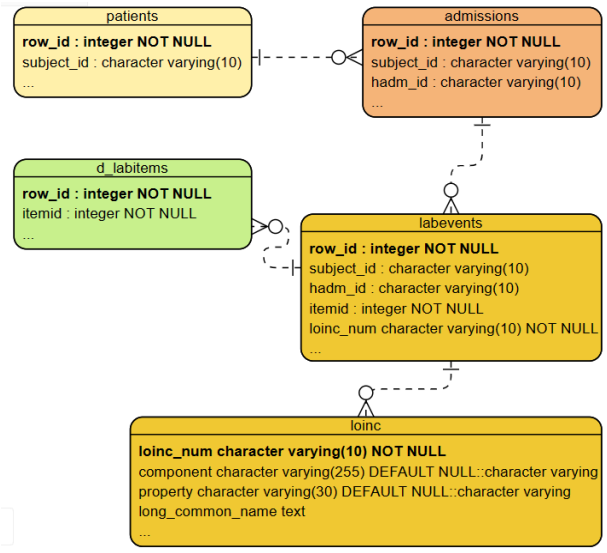

Figura 4: estrutura parcial do MIMIC III Postgres SQL

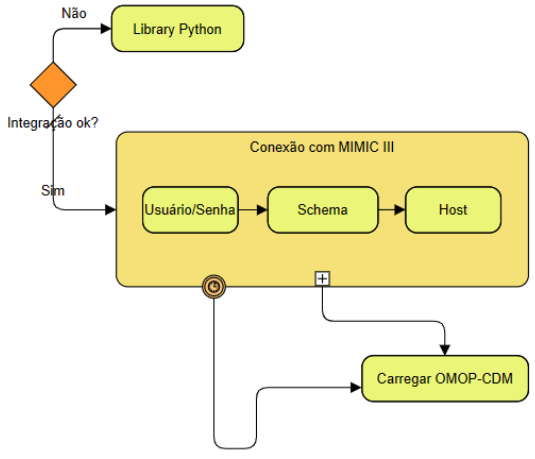

Figura 5: representação utilizada no ambiente de desenvolvimento Microsoft Visual Code.

O repositório BigQuery facilitou a execução de diversas etapas de pré-processamento, incluindo a remoção de duplicações e a correção de inconsistências, além da padronização dos campos para garantir a uniformidade e a integridade dos dados. Durante a integração dos dados no modelo OMOP-CDM, enfrentou-se o desafio de categorizar informações não estruturadas nas notas clínicas dos pacientes, especialmente as relacionadas a tabagismo, etilismo e obesidade. Todos os dados foram anonimizados antes do carregamento no ambiente BigQuery, usando técnicas de pseudo-anonimização para manter a privacidade dos pacientes e assegurar a reversibilidade caso necessário.

5.5. Categorização de Dados Não Estruturados

Um aspecto crítico do trabalho foi a estruturação dos dados não estruturados relacionados ao tabagismo, etilismo e obesidade. Inicialmente, foram criadas regras básicas de categorização para os dados textuais, que foram posteriormente aprimoradas para lidar com formas variadas de negação e expressões relacionadas. A categorização foi refinada para incluir uma coluna adicional, CATEGORY_2, que incorporou regras mais detalhadas e uma gama mais ampla de termos e expressões. Essas regras foram projetadas para melhorar a precisão da classificação, levando em conta diferentes formas de negação e terminologia relacionada ao tabagismo, etilismo e obesidade.

```sql
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
);
```

Figura 6: Regra utilizada para categorizar os dados.

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

Após a categorização dos dados não estruturados, a próxima etapa foi a aplicação de técnicas de aprendizado de máquina para a classificação de pacientes fumantes, a partir das descrições textuais extraídas dos prontuários médicos. O objetivo era identificar automaticamente menções ao tabagismo nos textos, utilizando modelos de classificação treinados em dados processados.

```sql
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;
```

Figura 7: Consulta utilizada binarizar categorias.

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

Figura 8: Exemplo de target criado.

5.6.1. Pré-processamento Textual

O pré-processamento textual é uma etapa fundamental para preparar os dados textuais de forma que possam ser adequadamente interpretados e utilizados pelos modelos de aprendizado de máquina. Cada etapa do pré-processamento tem como objetivo transformar os textos em estruturas numéricas ou vetoriais que representem o conteúdo semântico e sintático das palavras e frases. Esse processo inclui diversas operações, como tokenização, remoção de stopwords, normalização e vetorização, cada uma desempenhando um papel crucial na estruturação dos dados para a fase de modelagem.

5.6.1.1. Tokenização

A tokenização é o primeiro passo no pré-processamento e consiste na divisão do texto bruto em unidades menores, chamadas de tokens. Cada token representa uma palavra ou símbolo, e essa etapa é essencial para facilitar a

análise e manipulação subsequente dos dados. Por exemplo, dada a frase: "O paciente fuma dois maços de cigarro por dia." A tokenização dividiria o texto nas seguintes unidades: ["O", "paciente", "fuma", "dois", "maços", "de", "cigarro", "por", "dia"]. Esses tokens permitem que os modelos de aprendizado de máquina tratem o texto em segmentos discretos, facilitando o processamento e análise subsequente.

5.6.1.2. Remoção de Stopwords

Após a tokenização, é realizada a remoção das chamadas stopwords, que são palavras comuns e de baixa relevância para a classificação ou análise semântica, como "o", "de", "por", etc. Ao eliminar essas palavras, os modelos podem focar em termos mais relevantes que contribuem para a compreensão do contexto e do significado do texto. Por exemplo, após remover as stopwords da frase *tokenizada* anterior, restariam apenas as palavras de interesse: ["paciente", "fuma", "dois", "maços", "cigarro", "dia"]. Com a remoção de palavras irrelevantes, a análise passa a se concentrar nos termos essenciais para a tarefa de classificação.

5.6.1.3. Normalização

A normalização visa padronizar o texto para evitar inconsistências que possam surgir devido a variações de capitalização e acentuação. Nessa etapa, todo o texto é convertido para letras minúsculas e os acentos são removidos. Essa padronização é importante para garantir que palavras semanticamente iguais, mas escritas de forma diferente, sejam tratadas de forma consistente. Por exemplo:

Original: "Fuma dois Maços de Cigarro por Dia"
Normalizado: "fuma dois macos de cigarro por dia"

Esse processo evita que o modelo interprete "Fuma" e "fuma" como palavras diferentes, garantindo uma maior coerência nas representações.

5.6.1.4. Vetorização

Uma vez que os textos foram tokenizados, limpos e normalizados, eles precisam ser convertidos em representações numéricas para que possam ser processados pelos algoritmos de aprendizado de máquina. Para isso, são usadas técnicas de vetorização, que transformam os tokens em vetores numéricos que representam a frequência ou o contexto semântico das palavras.

Duas abordagens principais foram utilizadas neste trabalho: TF-IDF (Term

Frequency-Inverse Document Frequency) e embeddings gerados por modelos de linguagem (LLMs).

5.6.2. Representação Vetorial

Após o pré-processamento textual, é necessário transformar os textos em representações numéricas que possam ser interpretadas pelos algoritmos de aprendizado de máquina. Essa etapa de vetorização converte as palavras ou frases em vetores numéricos, permitindo que os modelos processem e analisem os textos de maneira eficiente. Existem diversas técnicas para realizar essa conversão, sendo as mais comuns TF-IDF e embeddings de modelos de linguagem (LLMs).

5.6.2.1. TF-IDF (Term Frequency-Inverse Document Frequency)

O TF-IDF é uma técnica que transforma cada palavra em um valor numérico que representa sua importância em um documento específico dentro de um conjunto de textos. O Term Frequency (TF) calcula quantas vezes uma palavra aparece em um documento, enquanto o Inverse Document Frequency (IDF) ajusta esse valor considerando a raridade da palavra em todo o corpus. O resultado é um vetor esparso, onde cada entrada corresponde à importância relativa de uma palavra em um documento.

Por exemplo, na frase: "O paciente fuma dois maços de cigarro por dia." Se a palavra "fuma" aparecer com frequência em muitos documentos, enquanto "cigarro" for menos comum, o TF-IDF atribui um valor maior a "cigarro", indicando sua maior relevância no contexto daquele documento específico. A representação da frase por um vetor TF-IDF poderia ser: [0.0, 0.0, 0.45, 0.0, 0.2, 0.35, 0.15, ...] onde cada posição no vetor corresponde a uma palavra do vocabulário, e os valores indicam a importância de cada termo no documento em questão.

5.6.2.2. Embeddings de Modelos de Linguagem (LLMs)

Diferente do TF-IDF, que se baseia na frequência de palavras, os embeddings gerados por Modelos de Linguagem (LLMs), como Word2Vec, BERT ou GPT, capturam relações semânticas mais profundas entre palavras e frases, considerando o contexto em que essas palavras aparecem. Esses embeddings são vetores densos, onde cada dimensão reflete uma característica semântica da palavra ou frase.

Por exemplo, a palavra "cigarro" pode ser representada por um vetor denso como: [0.12, -0.03, 0.45, 0.22, -0.35, 0.67, ...]. Palavras com significados semelhantes, como "fumo" ou "tabagismo", teriam vetores próximos no espaço vetorial, refletindo a similaridade de seus significados, mesmo que apareçam em diferentes contextos. Além disso, uma frase inteira pode ser representada por um único vetor que capte seu significado semântico geral no contexto do documento: [0.18, -0.24, 0.67, -0.05, 0.89, -0.15, ...]. Ao utilizar embeddings, o modelo de aprendizado de máquina não apenas processa a frequência de palavras, mas também é capaz de capturar o significado semântico e o contexto em que as palavras aparecem, proporcionando uma compreensão mais rica e profunda do conteúdo textual.

Assim, a vetorização, seja através de TF-IDF ou de embeddings de LLMs, é uma etapa crucial para garantir que os textos pré-processados sejam transformados em representações computacionais adequadas para os modelos de aprendizado de máquina, melhorando a qualidade e a eficiência das previsões.

5.6.3. Modelos de Classificação

Para a tarefa de classificação, dois modelos supervisionados de aprendizado de máquina foram selecionados: Regressão Logística e Árvore de Decisão.

5.6.3.1. Regressão Logística

A Regressão Logística foi escolhida por sua eficiência e simplicidade em problemas de classificação binária, onde o objetivo é categorizar instâncias em uma de duas possíveis classes. Este modelo tem a vantagem de fornecer resultados interpretáveis por meio de coeficientes que indicam a contribuição de cada variável preditora na probabilidade de ocorrência de uma determinada classe. No contexto deste estudo, ele se mostrou adequado para tarefas como identificar se um paciente é tabagista ou não, obeso ou não, ou etilista ou não, com base em representações textuais de dados clínicos.

Além disso, a Regressão Logística foi treinada com representações textuais geradas por TF-IDF e embeddings de Modelos de Linguagem (LLMs), que permitiram capturar nuances complexas presentes nos registros médicos. Por exemplo, expressões como "hipotireoidismo subclínico #tabagismo ativo" ou "nega tabagismo" foram devidamente processadas para alimentar o modelo. A escolha desse algoritmo também considerou sua eficiência computacional, especialmente relevante para a análise de grandes volumes de dados textuais em escala

multicêntrica, e sua capacidade de generalização em tarefas onde os dados são bem representados por relações lineares.

5.6.3.2. Árvore de Decisão

A Árvore de Decisão foi selecionada por sua capacidade de lidar com dados multivariados e de capturar interações não lineares entre variáveis. No contexto clínico, os dados são frequentemente heterogêneos, apresentando variações significativas devido a diferenças regionais, práticas hospitalares, estilos de documentação e até mesmo terminologias específicas de cada profissional de saúde. Essa variabilidade é ampliada em um estudo multicêntrico, onde os dados provêm de múltiplos hospitais, regiões e países, cada qual com suas particularidades.

A Árvore de Decisão é particularmente eficaz nesse tipo de cenário devido à sua natureza hierárquica, que permite segmentar dados de forma iterativa com base em características relevantes, mesmo quando essas características apresentam alta variabilidade. Além disso, o modelo é inerentemente interpretável, facilitando a validação clínica de suas decisões por especialistas da área de saúde.

No presente estudo, a Árvore de Decisão foi treinada utilizando as mesmas representações textuais (TF-IDF e embeddings), explorando a capacidade do modelo de identificar padrões latentes em dados textuais que podem não ser capturados por modelos lineares. Essa escolha foi motivada pela necessidade de construir um modelo robusto que pudesse lidar com a ambiguidade e a diversidade inerentes aos dados clínicos em um ambiente multicêntrico, onde os padrões são muitas vezes inconsistentes e influenciados por múltiplos fatores.

5.6.3.3. Implementação e Treinamento dos Modelos

A implementação e o treinamento dos modelos foram realizados em Python, utilizando a biblioteca scikit-learn. O processo foi dividido nas seguintes etapas:

- Divisão dos Dados: Os dados foram divididos em 80% para treinamento e 20% para teste, utilizando a função train_test_split para garantir a representatividade da amostra de teste.
- Treinamento dos Modelos: Ambos os modelos, Regressão Logística e Árvore de Decisão, foram treinados separadamente com representações TF-IDF e embeddings dos textos.

Trecho de código (python):

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.tree import DecisionTreeClassifier

# Regressão Logística
modelo_logistico = LogisticRegression()
modelo_logistico.fit(X_train_tfidf, y_train)

# Árvore de Decisão
modelo_arvore = DecisionTreeClassifier()
modelo_arvore.fit(X_train_tfidf, y_train)
```

- Balanceamento de Classes com SMOTE: O SMOTE (Synthetic Minority Over-sampling Technique) foi utilizado para gerar exemplos sintéticos da classe minoritária (não fumante), balanceando as classes e melhorando a acurácia em cenários desbalanceados.
- Otimização com Grid Search: O hiperparâmetro C da Regressão Logística foi ajustado com Grid Search e validação cruzada. O melhor valor encontrado foi C=1 com penalidade L2 e solver liblinear.

5.6.3.4. Avaliação dos Modelos

As métricas de avaliação incluíram acurácia, precisão, recall, F1-score e matriz de confusão. O código de avaliação é apresentado a seguir:

Trecho de código (python):

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

# Predições
y_pred_logistico = modelo_logistico.predict(X_test_tfidf)
y_pred_arvore = modelo_arvore.predict(X_test_tfidf)

# Avaliação Regressão Logística
acuracia_logistica = accuracy_score(y_test, y_pred_logistico)
precision_logistica = precision_score(y_test, y_pred_logistico)
recall_logistica = recall_score(y_test, y_pred_logistico)
f1_logistica = f1_score(y_test, y_pred_logistico)
matriz_confusao_logistica = confusion_matrix(y_test, y_pred_logistico)

# Avaliação Árvore de Decisão
```

```
        acuracia_arvore = accuracy_score(y_test, y_pred_arvore)
        precision_arvore = precision_score(y_test, y_pred_arvore)
        recall_arvore = recall_score(y_test, y_pred_arvore)
        f1_arvore = f1_score(y_test, y_pred_arvore)
        matriz_confusao_arvore = confusion_matrix(y_test,
y_pred_arvore)
```

6. RESULTADOS

Um dos principais desafios enfrentados durante a integração de dados no modelo OMOP-CDM foi o tratamento de informações não estruturadas presentes nas notas clínicas dos pacientes, especialmente aquelas relacionadas a tabagismo, etilismo e obesidade. Essas notas, frequentemente escritas em linguagem natural pelos profissionais de saúde, continham informações críticas, mas eram difíceis de categorizar de forma consistente para uso em modelos preditivos devido à falta de estrutura.

6.1. Importância da Integração de Dados não Estruturados nas Predições Clínicas

A integração de dados clínicos não estruturados em modelos preditivos representa uma abordagem essencial para potencializar a análise preditiva em saúde. Embora os registros eletrônicos de saúde (RES) tenham sido amplamente adotados, a maioria das pesquisas científicas ainda foca em dados estruturados — como diagnósticos e exames laboratoriais — que são mais facilmente quantificáveis e categorizados. Em contrapartida, muitos dados clínicos críticos, especialmente informações sobre hábitos de vida, como histórico de tabagismo, etilismo e obesidade, são registrados em campos de texto livre nas notas clínicas. Essas descrições, por capturarem nuances das condições e comportamentos dos pacientes, possuem grande valor para a compreensão do perfil clínico e dos fatores de risco, mas frequentemente são subutilizadas em modelos de predição devido à complexidade envolvida em extrair e categorizar essas informações.

De acordo com Spasic e Nenadic (2020), "o processamento de linguagem natural (PNL) demonstrou repetidamente sua viabilidade para desbloquear evidências enterradas em narrativas clínicas". Isso sublinha o potencial das técnicas de PNL em aproveitar esses dados textuais, transformando-os em variáveis estruturadas que possam ser incorporadas de forma fluida aos modelos preditivos. A ausência de estrutura e a variabilidade no vocabulário utilizado por profissionais de saúde dificultam o uso direto desses dados, muitas vezes resultando na exclusão de informações valiosas que poderiam enriquecer significativamente a precisão e a profundidade das predições de eventos cardiovasculares adversos maiores (MACE).

Este estudo busca explorar abordagens avançadas de categorização de

dados não estruturados, aplicando modelos de linguagem de grande escala (LLMs) para processar e organizar as informações de texto livre. O foco está na padronização e na categorização de descrições relacionadas ao tabagismo, etilismo e obesidade, variáveis de alto impacto para as predições de MACE. No tópico 6.2, é detalhado o processo específico de categorização do tabagismo, uma das principais variáveis de risco analisadas neste estudo.

6.2. Processamento Inicial e Categorização de Tabagismo

Inicialmente, foi utilizado um método de categorização baseado na presença de palavras-chave em campos de texto livre, como “tabagismo” e “tabagista”. Este processo, embora eficiente para capturar ocorrências diretas, apresentou limitações ao lidar com negações complexas ou variações semânticas. Expressões como "cessou tabagismo" ou "ex-tabagista" não foram corretamente classificadas, evidenciando a necessidade de uma abordagem mais robusta.

Para resolver essa limitação, a categorização foi refinada por meio de regras mais sofisticadas que incorporaram um conjunto maior de expressões. Isso resultou na criação de uma nova coluna, CATEGORY_2, dentro da tabela Smoking_Processed_Combined. Este refinamento incluiu o uso de SQL avançado no BigQuery, utilizando funções como STRING_AGG e CASE WHEN, para categorizar de forma precisa variações de linguagem que indicavam tanto o histórico quanto o status atual de tabagismo. Foram incluídas condições que capturavam expressões variadas como "%cessou tabagismo%", "%ex-tabagista%", e "%negado etilismo ou tabagismo%". A análise temporal e de contexto também foi implementada, diferenciando, por exemplo, frases que indicavam que o paciente havia parado de fumar no passado de outras que sugeriam que a orientação para cessação ainda estava em curso.

6.3. Impacto da Categorização no Machine Learning

A categorização aprimorada permitiu transformar dados textuais complexos em um target binário (‘smoke’ e ‘nosmoke’), fundamental para o treinamento de modelos de aprendizado de máquina. Esse processo foi crucial para a criação de predições mais precisas de MACE, já que o tabagismo é um dos principais fatores de risco na avaliação de eventos cardiovasculares. Este avanço não apenas melhorou a qualidade dos dados textuais utilizados, mas também demonstrou o impacto direto do pré-processamento e da categorização refinada na eficácia dos

modelos de machine learning aplicados em dados clínicos não estruturados.

6.4. Aplicação de Modelos de Machine Learning

Para automatizar a classificação de tabagismo, dois modelos de aprendizado de máquina supervisionados foram aplicados: Regressão Logística e Árvores de Decisão.

6.4.1. Regressão Logística

Utilizada pela sua simplicidade e capacidade de resolver problemas de classificação binária, a Regressão Logística foi treinada com vetores TF-IDF e embeddings de LLMs, alcançando uma acurácia de 91,6%. Esse modelo foi eficaz em associar palavras-chave relacionadas ao tabagismo à classificação final de fumante ou não fumante, sendo adequado para estudos multicêntricos, onde a padronização e a simplicidade na interpretação são essenciais.

6.4.2. Árvores de Decisão

Para capturar interações mais complexas nos textos, o modelo de Árvores de Decisão foi empregado, permitindo segmentar informações textuais de maneira não linear. Embora a acurácia (89,4%) tenha sido ligeiramente inferior à da regressão logística, a capacidade interpretativa do modelo e sua flexibilidade ao lidar com descrições indiretas ou ambíguas tornam-no valioso para a análise clínica.

6.5. Avaliação dos Modelos e Métricas de Desempenho

Os modelos foram avaliados utilizando métricas como acurácia, precisão e recall, sendo a matriz de confusão a principal ferramenta de análise. Conforme demonstrado na Tabela 1, 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 1: Matriz de Confusão da Regressão Logística

Tabela 1: Matriz de Confusão da Regressão Logística	
Previsão: Fumante	Previsão: Não Fumante
Fumante: 94	6

Não Fumante: 5	95

A regressão logística alcançou uma acurácia de 91,6%, enquanto a árvore de decisão apresentou uma acurácia de 89,4%, conforme resumido na Tabela 2.

Tabela 2: Desempenho dos Modelos

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

6.6. Análise Comparativa: TF-IDF e Embeddings LLM

Uma contribuição central deste estudo foi a comparação entre abordagens clássicas, como o TF-IDF, e métodos mais modernos, baseados em embeddings de LLMs. O uso dessas técnicas permitiu capturar diferentes nuances semânticas dos textos, especialmente em relação à complexidade das descrições clínicas de tabagismo. A combinação dessas abordagens não apenas aumentou a acurácia dos modelos, mas também permitiu uma maior flexibilidade e robustez ao lidar com variações linguísticas nos registros clínicos.

Considerando os resultados desses modelos pensando nas implicações para a Predição de MACE, ambos os modelos demonstraram sua eficácia na classificação de tabagismo, uma variável crítica para a predição de MACE. A regressão logística, com sua simplicidade e alta acurácia, mostrou-se ideal para estudos multicêntricos de grande escala. Por outro lado, a árvore de decisão, com sua capacidade de modelar interações mais complexas, é particularmente útil para análises detalhadas que exigem a compreensão de nuances textuais em registros médicos. Esses resultados reforçam a importância de técnicas avançadas de processamento de linguagem natural e sua aplicação em modelos preditivos na área da saúde, enfatizando a necessidade contínua de integrar dados não estruturados de forma eficiente para melhorar a precisão e a usabilidade dos modelos de aprendizado de máquina.

Na primeira etapa, o texto na coluna CTU_INFORMACAO é pré-processado de forma mínima, priorizando a coerência semântica, sem aplicar transformações excessivas. Em seguida foi utilizado o modelo *paraphrase-multilingual-MiniLM-L12-v2* da biblioteca *SentenceTransformer*, que é eficaz em várias línguas, para gerar embeddings numéricos de alta dimensionalidade (384 dimensões) para cada frase ou sentença. O objetivo aqui é transformar cada texto em um vetor que capture suas características semânticas para alimentar modelos baseados em aprendizado profundo. Estes embeddings são extraídos a partir do texto original para preservar a riqueza semântica, fornecendo ao modelo representações numéricas que facilitam o entendimento do conteúdo do texto.

Os embeddings extraídos são armazenados em um DataFrame denominado dfe, no qual cada linha representa uma frase convertida em um vetor de 384 componentes. Abaixo está uma amostra dos primeiros vetores gerados:

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

Tabela 3: Vetores gerados

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 são então utilizados como entrada em modelos de aprendizado de máquina, como no caso da predição de eventos cardiovasculares adversos maiores (MACE). A principal vantagem de utilizar embeddings reside na sua capacidade de capturar a semântica subjacente do texto, facilitando a identificação de padrões relevantes para a tarefa em questão.

Ademais, foi aplicada uma abordagem complementar com o uso da técnica TF-IDF (Term Frequency-Inverse Document Frequency). Diferente dos embeddings, o TF-IDF gera representações numéricas focadas na frequência e

importância dos termos, utilizando um pré-processamento mais rigoroso que envolve a remoção de stopwords, lematização e normalização do texto. Isso resulta em vetores esparsos que priorizam os termos mais representativos do conteúdo textual. Dessa forma, essa abordagem híbrida—ao combinar embeddings para capturar a semântica e TF-IDF para capturar a importância e frequência das palavras—oferece um modelo robusto, capaz de explorar diferentes dimensões dos dados textuais. Assim, essa estratégia melhora a precisão do modelo na predição de MACE, explorando tanto a riqueza semântica quanto às frequências textuais.

7. Estudo dos Resultados dos Modelos

Os modelos gerados a partir dessas representações foram avaliados utilizando métricas de desempenho em conjuntos de treino e teste, permitindo a análise da capacidade de generalização dos modelos. A análise inicial dos resultados, representada pela curva ROC na Figura 9, revelou que o modelo treinado com as representações textuais apresentava uma área sob a curva (AUC) de 0.68 no conjunto de treino, sugerindo uma capacidade moderada de discriminar entre pacientes que poderiam ou não experienciar MACE. No entanto, ao avaliar o desempenho no conjunto de teste, o AUC caiu para 0.47, conforme visto na Figura 10, indicando que o modelo se aproximava de um comportamento aleatório na classificação dos casos. Essa diferença de desempenho entre treino e teste sugere um possível problema de sobreajuste, onde o modelo aprendeu características específicas do conjunto de treino que não se generalizam bem para novos dados.

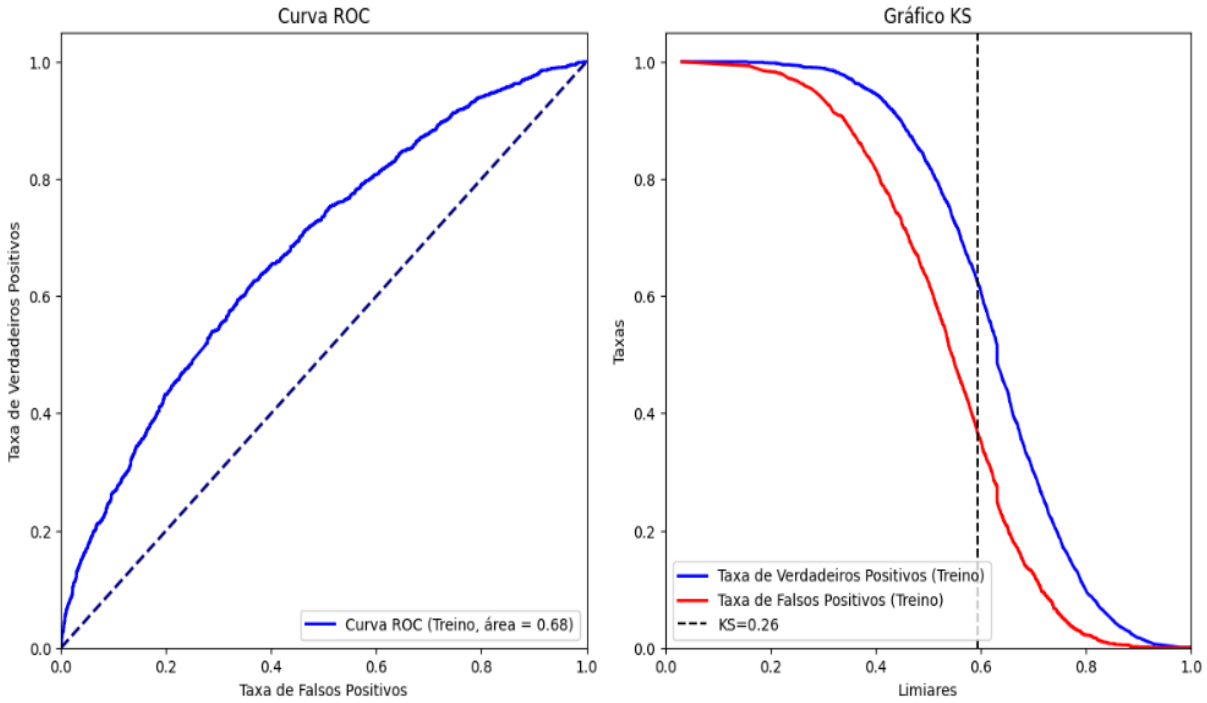

Figura 9: Curva ROC das classificações textuais do Tabagismo.

Além disso, a análise de separação entre as classes foi complementada pelo gráfico de Kolmogorov-Smirnov (KS), ilustrado na Figura 10. No conjunto de treino, a métrica KS foi de 0.26, evidenciando uma separação moderada entre as distribuições das classes. Contudo, o desempenho no conjunto de teste foi dramaticamente inferior, com um KS de apenas 0.01, o que confirma a baixa capacidade do modelo em discriminar entre pacientes com e sem MACE na coorte estudada, considerando somente o atributo “tabagismo”, ao ser exposto a novos dados.

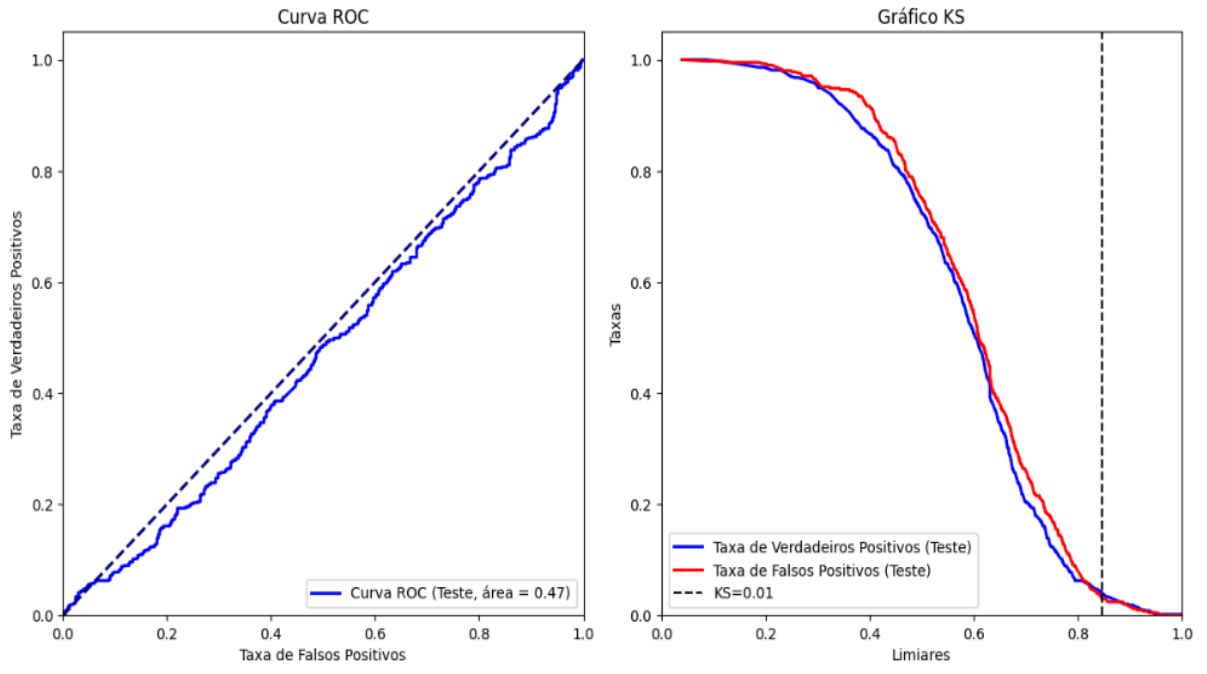

Figura 10: Desempenho da separação entre as classes.

Outra métrica fundamental para essa avaliação foi a matriz de confusão, mostrada na Figura 11, que expôs a fragilidade do modelo no conjunto de teste. Embora o modelo tenha conseguido identificar corretamente 338 instâncias da classe negativa, ele apresentou dificuldades em classificar corretamente os eventos positivos (MACE), com um elevado número de falsos negativos (518 casos). A baixa performance nesse aspecto é preocupante, visto que o objetivo principal é justamente identificar corretamente os casos positivos, para que medidas preventivas possam ser tomadas a tempo.

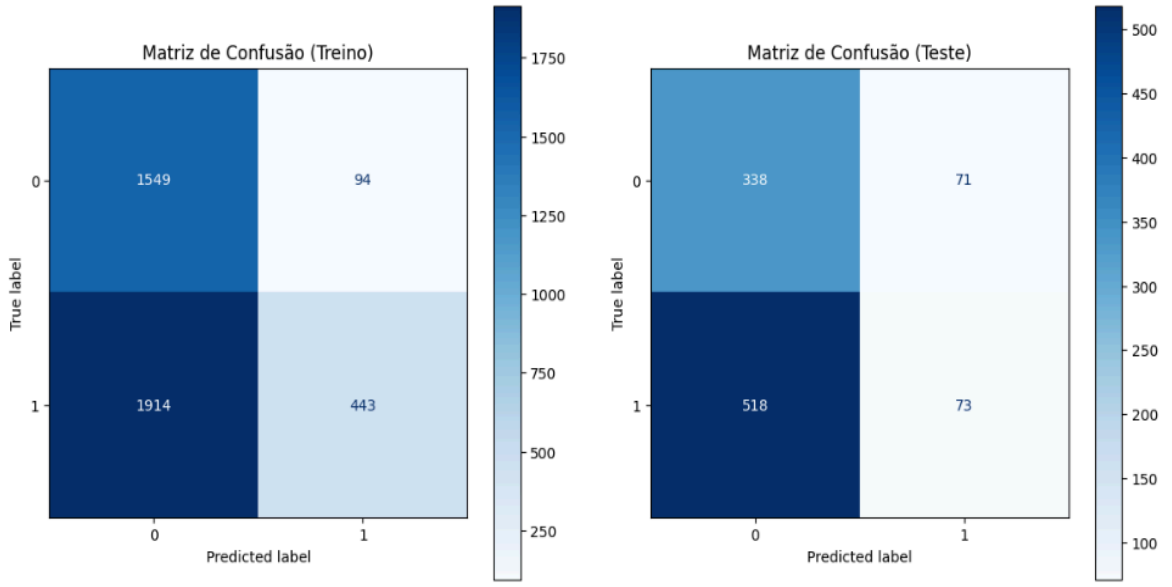

Figura 11: Matriz de confusão das classes.

Diante desse cenário, foi necessário aplicar técnicas de balanceamento das classes para melhorar a capacidade do modelo em identificar os casos positivos. Após o balanceamento o desempenho foi reavaliado e houve uma melhora significativa nas métricas de classificação, principalmente no que tange à classe 1 (casos positivos de MACE). A precisão (precision) da classe 1 subiu para 0.59, indicando uma redução nos falsos positivos, enquanto o recall da mesma classe alcançou 0.52, demonstrando que o modelo passou a capturar uma fração maior dos eventos de MACE. Em contrapartida, a precisão para a classe negativa (0.40) ainda reflete uma taxa de falsos positivos relativamente alta, o que pode ser um ponto de atenção, dependendo das implicações práticas dessa aplicação.

Tabela 4: Desempenho do modelo sem utilizar os novos campos estruturados.

Matriz de Confusão	
193	216
284	307

A análise das matrizes de confusão após o balanceamento, representada na Tabela 4, confirma que o modelo passou a identificar melhor a classe positiva. No entanto, o trade-off entre a melhoria na detecção de MACE e a queda na performance da classe negativa sugere que ainda é necessário ajustar o modelo para evitar penalizar demais uma classe em favor da outra. O F1-score equilibrado de 0.44 para a classe negativa e 0.55 para a positiva reflete essa situação, onde o modelo consegue capturar melhor a classe relevante (MACE), mas com uma acurácia global de apenas 50%.

Esses resultados, tanto antes quanto após o balanceamento, indicam que o modelo se beneficia da combinação de embeddings e TF-IDF na representação dos textos, mas ainda carece de ajustes mais refinados para obter uma performance adequada nos dois cenários (treino e teste). O uso de técnicas de regularização, ajuste de hiperparâmetros ou a inclusão de variáveis clínicas adicionais pode melhorar a generalização do modelo e, consequentemente, sua eficácia na predição de eventos cardiovasculares adversos.

7. DISCUSSÃO

Melhorar a eficácia de predições e generalizações para diferentes populações é significativo com dados bem estruturados e classificados, pois estes aumentam significativamente a capacidade dos modelos de capturar padrões de risco relevantes. Esse aspecto torna-se especialmente importante ao se lidar com eventos adversos multifatoriais, como os MACE, que exigem uma análise ampla de variáveis. A inclusão de variáveis adicionais, como etilismo e obesidade, em formato estruturado demonstrou que esses fatores enriquecem a análise e ampliam a precisão do modelo, refletindo a complexidade dos fatores de risco envolvidos em eventos cardiovasculares.

A integração das bases do HCFMRP e MIMIC-III trouxe um ganho substancial, pois possibilitou a exposição dos modelos a uma ampla gama de dados demográficos e clínicos, aprimorando a capacidade dos modelos de AM de generalizar para contextos clínicos variados. A padronização promovida pelo OMOP-CDM facilitou essa integração, promovendo uma colaboração eficaz entre instituições e permitindo a realização de estudos multicêntricos. Com essa diversidade de dados, os modelos foram capazes de lidar melhor com diferentes perfis populacionais, o que aumenta suas chances de aplicação prática e relevância clínica.

Uma das maiores dificuldades encontradas foi a inclusão de dados textuais não estruturados, particularmente relacionados à tabagismo, etilismo e obesidade, que são informações clínicas frequentemente registradas de forma variada em campos de texto livre. Esses campos, que variam em terminologia, estilo e contexto, exigem uma abordagem de processamento específica para capturar nuances sem perder a precisão. A estruturação de campos de texto livre foi, assim, uma das etapas mais desafiadoras e inovadoras deste projeto, devido à complexidade e variabilidade inerente a esses dados.

No contexto do PRE-CARE ML, cuja meta é validar modelos preditivos em ambientes clínicos diversos e criar métricas generalizáveis, a inclusão de atributos no modelo depende de sua disponibilidade em todos os multicentros participantes. Informações como tabagismo, etilismo e obesidade, registradas de forma textual e informal nos prontuários médicos do HCFMRP, contrastaram com outros hospitais onde dados semelhantes já estavam previamente categorizados em campos estruturados. Essa discrepância resultou na impossibilidade de integrar esses

fatores de risco de maneira uniforme, gerando uma perda significativa de informações potencialmente relevantes para o modelo.

A tabela Smoking_Processed_Combined reflete o esforço para estruturar essas informações no HCFMRP, transformando dados textuais livres em variáveis analisáveis. Contudo, a ausência de padronização completa entre os multicentros reforça a necessidade de uniformização de registros clínicos estruturados. Essa uniformização é essencial para garantir que variáveis clínicas críticas, como tabagismo e obesidade, possam ser integradas a análises multicêntricas, maximizando a representatividade e aplicabilidade clínica dos modelos preditivos.

Para lidar com esses desafios, utilizamos uma combinação de técnicas de processamento de linguagem natural (PLN) que permitiram transformar as descrições textuais em variáveis categorizadas e estruturadas. O processo de estruturação dos dados textuais envolveu várias etapas, começando pela tokenização e normalização dos registros. Cada entrada textual foi tokenizada para identificar palavras e expressões-chave, enquanto a normalização incluiu a remoção de pontuações e a transformação de letras para minúsculas. Esse processo de normalização foi essencial para uniformizar termos registrados de forma variável, como “fumante,” “fuma,” “não fuma,” “parou de fumar,” e “ex-fumante,” facilitando uma análise mais precisa e consistente.

A seguir, foram aplicadas regras de identificação de palavras-chave e frases específicas, utilizando listas de termos relacionados a cada condição. No caso de tabagismo, por exemplo, palavras e expressões como “fumante,” “ex-fumante,” “parou de fumar,” e “nunca fumou” foram selecionadas e extraídas. Essa etapa inicial baseou-se em métodos de reconhecimento de padrões, com o objetivo de capturar o máximo de ocorrências relevantes, assegurando a cobertura abrangente dos registros.

Um dos principais desafios enfrentados foi a interpretação correta de negações e tempos verbais, que alteram significativamente o significado das informações textuais. Para resolver esse problema, foi desenvolvida uma lógica específica para reconhecer negações como “não fuma,” “nunca fumou,” e “nega tabagismo", que incluía regras de detecção de negações e análise gramatical do contexto. Esse processo foi fundamental para evitar classificações incorretas, assegurando que o contexto semântico de cada expressão fosse compreendido e processado adequadamente. A fim de capturar a semântica subjacente e melhorar a interpretação de frases complexas, foram empregados embeddings de linguagem

gerados por modelos de linguagem de larga escala (LLMs). Essa técnica permitiu que o modelo detectasse padrões e relações semânticas, mesmo quando a estrutura das frases ou a terminologia variavam, o que foi especialmente útil para distinguir condições como “fumante ativo” e “ex-fumante,” além de lidar com sinônimos e expressões coloquiais.

Após o pré-processamento, os registros foram classificados em categorias estruturadas, como “fumante ativo,” “ex-fumante,” “não fumante,” e “status desconhecido.” Essas categorias foram convertidas em variáveis binárias ou categóricas, tornando os dados prontos para uso direto nos modelos de aprendizado de máquina (AM). A tabela Smoking_Processed_Combined exemplifica essa estruturação, ao converter informações textuais não estruturadas em um formato que pode ser utilizado eficazmente pelos modelos preditivos.

Para assegurar a precisão da categorização, realizamos uma validação manual em uma amostra dos dados, o que nos permitiu identificar ambiguidades e ajustar as regras de categorização. Também testamos a consistência dos embeddings para verificar se as relações semânticas capturadas eram apropriadas ao contexto clínico. No entanto, percebemos a necessidade de realizar mais testes de consistência com os embeddings para refinar ainda mais o modelo. Devido a limitações de tempo, não foi possível conduzir um número maior desses testes, ficando como uma importante sugestão para trabalhos futuros a ampliação dessa etapa de validação.

Embora as métricas de desempenho, como acurácia e precisão, não tenham alcançado valores elevados, o trabalho demonstrou ser altamente promissor em termos de escalabilidade e potencial de evolução. A transformação eficaz de campos de texto livre em variáveis estruturadas permitiu integrar fatores de risco adicionais nos modelos de predição de MACE, o que, apesar das métricas ainda moderadas, representa um avanço significativo ao enriquecer o modelo com informações previamente subutilizadas. Essa abordagem de estruturação de dados textuais não só viabiliza a incorporação de outros fatores documentados em registros clínicos, como padrões alimentares, práticas de atividade física e condições médicas autodeclaradas, mas também abre caminho para que os modelos se tornem cada vez mais abrangentes e precisos ao identificar riscos cardiovasculares à medida que mais variáveis são integradas.

Concomitantemente, a metodologia também se mostrou particularmente promissora na classificação de tabagismo, um fator crucial para a predição de

MACE. A combinação de embeddings com TF-IDF permitiu capturar tanto a semântica dos textos quanto a relevância dos termos, aumentando a capacidade do modelo de identificar padrões associados ao risco cardiovascular, mesmo com a complexidade dos dados textuais.

Além disso, a aplicação de embeddings e técnicas de pré-processamento a essas variáveis textuais sugere uma possibilidade de replicação para outras informações não estruturadas dos prontuários médicos, como etilismo e obesidade, o que ampliaria ainda mais a capacidade dos modelos de estratificar riscos de forma precisa e personalizada. Essa abordagem escalável oferece grande potencial para a automação de análises em campos livres, representando um avanço para a personalização dos cuidados de saúde e a identificação específica de fatores de risco por paciente, o que se tornará mais robusto conforme novas variáveis forem incorporadas e refinadas em estudos futuros.

8. CONCLUSÃO

A conclusão desta monografia destaca os avanços e as contribuições alcançadas com o desenvolvimento de uma metodologia para estruturação de dados textuais em prontuários médicos, visando a predição de MACE (Eventos Cardiovasculares Adversos Maiores). O projeto possibilitou a transformação de informações originalmente não estruturadas, como tabagismo e outros fatores de risco, em variáveis estruturadas utilizáveis em modelos de aprendizado de máquina. Embora as métricas de desempenho, como acurácia e precisão, não tenham atingido os valores ideais, a abordagem mostrou-se promissora devido à sua escalabilidade e à possibilidade de expandir a análise para outras categorias de fatores de risco, como etilismo e obesidade, além de incluir mais variáveis ao longo do tempo.

As técnicas aplicadas, como o uso de embeddings combinados com TF-IDF, foram eficazes para capturar tanto a semântica quanto a relevância de termos nos registros textuais. Este processo permitiu uma melhor identificação de padrões e nuances nos textos, tornando o modelo mais sensível às variações semânticas que são fundamentais para a análise clínica. A abordagem híbrida para categorização de variáveis textuais abriu novas possibilidades de utilização dos registros médicos completos, permitindo análises mais ricas e contextualmente precisas, o que pode ser replicado para outros fatores de risco em projetos futuros.

Outro aspecto importante foi a capacidade de estruturar campos textuais livres de maneira automatizada, representando um avanço significativo para a personalização dos cuidados de saúde. Essa estruturação permite que modelos preditivos identifiquem fatores de risco específicos de cada paciente, contribuindo para uma estratificação de risco mais precisa e personalizada, essencial para intervenções clínicas preventivas e mais eficazes. Em continuidade, foi viável categorizar pacientes em relação ao tabagismo utilizando uma abordagem que integra aprendizado de máquina convencional com representações de texto geradas por modelos LLM.

Nessa metodologia, os LLMs tokenizam e transformam cada registro textual em vetores numéricos, permitindo ao classificador convencional identificar padrões semânticos e distinguir entre pacientes tabagistas e não tabagistas. Essa integração não apenas aprimora a precisão, como também demonstra a robustez do método, viabilizando sua replicação para outras variáveis

Em suma, a metodologia desenvolvida demonstra que, mesmo com métricas

ainda moderadas, o potencial de evolução e a escalabilidade da estruturação de dados textuais apontam para uma ampliação significativa nas capacidades dos modelos preditivos.

Nesse sentido, o trabalho oferece bons indicativos para estudos futuros, que poderão explorar novas variáveis e refinar as técnicas de categorização e análise semântica, contribuindo para o avanço da pesquisa em predição de MACE e personalização da saúde.

9. REFERÊNCIAS BIBLIOGRÁFICAS

Spasic, I., & Nenadic, G. (2020). Dados de texto clínico em aprendizado de máquina: revisão sistemática. Informática médica JMIR, 8(3), e17984. https://doi.org/10.2196/17984

Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., & Nichols, M. (2016). Cardiovascular disease in Europe. https://doi.org/10.1093/eurheartj/ehw334.

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.

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).

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://physionet.org/content/mimiciii/1.4/
Auxílio à pesquisa 21/06137-4 - Aprendizado computacional, Diagnóstico precoce - BV FAPESP. Disponível em: < https://bv.fapesp.br/pt/auxilios/110451/prevendo-eventos-cardiovasculares-usando-aprendizado-de-maquina />.

PreCare ML Project. Disponível em: < https://precareml.github.io />.

((Machine Learning) AND (health)) AND (predict) - Search Results - PubMed. Disponível em: < https://pubmed.ncbi.nlm.nih.gov/?term=%28%28Machine+Learning%29+AND+%28health%29%29+AND+%28predict%29 >. Acesso em: 29 ago. 2024.

Polat Erdeniz, S., Kramer, D., Schrempf, M., Rainer, P. P., Felfernig, A., Tran, T. N. T., Burgstaller, T., & Lubos, S. (2023). Machine Learning Based Risk Prediction for Major Adverse Cardiovascular Events for ELGA-Authorized Clinics1. Studies in health technology and informatics, 301, 20–25. https://doi.org/10.3233/SHTI230006

Shu, S., Ren, J., & Song, J. (2021). Aplicação clínica de inteligência artificial baseada em aprendizado de máquina no diagnóstico, previsão e classificação de doenças cardiovasculares. Jornal de circulação: jornal oficial da Sociedade de Circulação Japonesa, 85(9), 1416–1425. https://doi.org/10.1253/circj.CJ-20-1121

Krittanawong, C., Zhang, H. J., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 69(17), 2096-2107. https://doi.org/10.1016/j.jacc.2017.03.571