Universidade de São Paulo

Faculdade de Medicina de Ribeirão Preto

Departamento de Imagens Médicas, Hematologia e Oncologia Clínica

Divisão de Ciências das Imagens e Física Médica

Edital PUB-USP 2023

Estudo de organização de coorte para aprendizado de máquina utilizando o modelo OMOP-CDM da OHDSI

Mayara Martins Perroni

Coordenador: Prof. Dr. Paulo Mazzoncini de Azevedo Marques
Pesquisadores Colaboradores:
Bolsista FAPESP TT5 Hilton Vicente César

Ribeirão Preto – SP, Março 2024

Resumo. *O crescimento da tecnologia da informação em saúde abriu as portas para as organizações deste setor manterem seus dados em várias formas e em vários volumes. Tais dados têm sido usados para gerar conhecimento e apoiar a solução de vários problemas. A heterogeneidade da organização dos dados nas diferentes organizações tende a levar os modelos de aprendizado de máquina (AM) a overfitting local, ajustando-os somente aos detalhes específicos do conjunto de dados de treinamento, sem capacidade de generalização de seus resultados. Para dar suporte à análise de dados multicêntricos e heterogêneos, foram implementadas técnicas de Extraction-Transformation-Loading (ETL). Neste projeto, dados extraídos de registros eletrônicos de saúde (RES) foram padronizados utilizando-se o Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM), que é um padrão de dados da comunidade aberta Observational Health Data Sciences and Informatics (OHDSI).*

Abstract.*The growth of information technology in healthcare has enabled organizations in this sector to keep their data in various forms and volumes. These data have been used to generate knowledge and support the solution of several problems. The heterogeneity of data organization in different organizations tends to lead machine learning (ML) models to local overfitting, adapting them only to the specific details of training dataset, without the ability of generalizing their results. To support the analysis of multicentric and heterogeneous data, Extraction-Transformation-Loading (ETL) techniques have been implemented. In this project, data extracted from electronic health records (EHRs) were standardized using the Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM), which is a data standard from the open community Observational Health Data Sciences and Informatics (OHDSI).*

Palavras-chave: MACE; OMOP; ETL.

Palavras-chave: MACE; OMOP; CMD; OHDSI; ETL.

Nome do projeto: Validação e melhoria de modelos de aprendizado de máquina para previsão de eventos cardiovasculares.

1. INTRODUÇÃO

As doenças cardiovasculares (DCVs) representam a principal causa de mortalidade global, sendo responsáveis por milhões de óbitos anualmente. Entre as DCVs, a aterosclerose desempenha um papel central, levando ao desenvolvimento de condições críticas como infarto do miocárdio, doença cardíaca isquêmica e acidente vascular cerebral. Esses eventos não apenas contribuem significativamente para a morbidade e mortalidade, mas também impõem um enorme ônus econômico aos sistemas de saúde em todo o mundo. A identificação precoce de indivíduos com alto risco de eventos cardiovasculares adversos maiores (MACE - major adverse cardiovascular events) é essencial para a implementação de intervenções preventivas que podem mitigar o impacto dessas doenças.

Tradicionalmente, a estratificação de risco cardiovascular tem sido baseada em escores clínicos, como o ESC SCORE e o Framingham Risk Score, que utilizam um conjunto limitado de fatores de risco bem estabelecidos, como idade, pressão arterial, níveis de colesterol e tabagismo. No entanto, essas abordagens convencionais têm limitações, especialmente em termos de acurácia e capacidade de personalização. Recentemente, o aprendizado de máquina (ML - machine learning) emergiu como uma ferramenta promissora para melhorar a predição de risco, utilizando uma gama mais ampla de variáveis e permitindo a detecção de padrões complexos que podem escapar aos métodos tradicionais.

Dados extraídos recentemente do PubMed, em busca avançada (“((Machine Learning) AND (health)) AND (predict) - Search Results - PubMed”, 2024)), evidenciam um crescimento exponencial no número de publicações relacionadas ao uso de Machine Learning para predição na saúde ao longo dos últimos anos, destacando a crescente relevância dessa abordagem na medicina. Até 2014, o número de estudos era modesto, com menos de 20 publicações anuais, mas a partir de 2015 observou-se um aumento consistente que se intensificou significativamente nos anos subsequentes. Em 2024 já registra 4.674 publicações até o momento. O aumento do volume de dados de saúde, aliado aos avanços nas técnicas de Machine Learning, reforça o potencial transformador desta tecnologia na medicina moderna, apontando para um futuro em que as predições e intervenções clínicas sejam cada vez mais precisas, personalizadas e eficazes, superando as limitações dos métodos tradicionais.

Modelos de ML, como redes neurais profundas, árvores de decisão e métodos baseados em ensemble, têm demonstrado potencial para superar as abordagens tradicionais na predição de risco de MACE. Esses modelos são capazes de integrar múltiplos preditores de risco de maneira dinâmica e adaptativa, potencialmente fornecendo previsões mais precisas e personalizadas para os pacientes. No entanto, apesar do desenvolvimento acelerado de modelos de ML nos últimos anos, a validação desses modelos em contextos clínicos reais permanece um desafio. A maioria dos estudos concentra-se em coortes limitadas e homogêneas, o que levanta preocupações sobre a capacidade desses modelos de generalizar para populações diversas e sistemas de saúde variados.

Um dos maiores obstáculos para a generalização dos modelos de ML é a variabilidade inerente aos dados clínicos entre diferentes centros e populações. Essa variabilidade inclui diferenças nos sistemas de coleta de dados, nas práticas clínicas, nas características demográficas dos pacientes e nos recursos disponíveis em diferentes sistemas de saúde. Por exemplo, modelos desenvolvidos em populações europeias podem não apresentar o mesmo desempenho quando aplicados em contextos latino-americanos ou asiáticos, onde os perfis de

risco e as práticas clínicas podem diferir substancialmente. Essa heterogeneidade dos dados torna crítica a validação multicêntrica e multinacional de modelos de ML, para assegurar que suas previsões sejam robustas e aplicáveis a um amplo espectro de pacientes.

O presente estudo insere-se no contexto do projeto PRE-CARE ML: (“Auxílio à pesquisa 21/06137- 4 - Aprendizado computacional, Diagnóstico precoce - BV FAPESP”, 2024), que visa investigar a capacidade de generalização de modelos de aprendizado de máquina (ML) para a predição de eventos cardiovasculares adversos maiores (MACE) a partir de dados padronizados de registros eletrônicos de saúde (EHRs - electronic health records) de múltiplos centros. Esses modelos, desenvolvidos pela Medical University of Graz (MUG), na Áustria, e inicialmente testados em hospitais vinculados ao The Health Care Company of Styria (KAGes), serão validados e adaptados para diferentes cenários clínicos em colaboração com a Faculdade de Medicina de Ribeirão Preto – USP. Tal parceria permitirá explorar o desempenho desses modelos em populações diversificadas e heterogêneas, abrangendo pacientes de diferentes origens étnicas e com distintas prevalências de fatores de risco.

A validação de modelos de ML em um contexto multicêntrico e multinacional, incluindo hospitais na Áustria, Brasil, e possivelmente outros países, oferece um ganho científico significativo, especialmente no que diz respeito à robustez e à aplicabilidade desses modelos em diferentes populações. A diversidade nos dados clínicos, demográficos e nos padrões de saúde entre os centros é uma oportunidade para testar a capacidade de generalização dos modelos em cenários variados, refletindo a realidade de sistemas de saúde globais com perfis epidemiológicos distintos. Não apenas foi explorada a generalização dos modelos, mas também o avanço das estratégias de predição de risco cardiovascular, fornecendo um conjunto de evidências robustas sobre o uso de ML em diferentes contextos clínicos.

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo geral da bolsa de iniciação científica foi preparar e estruturar dados provenientes de múltiplas fontes para permitir a sua integração e utilização efetiva em modelos de aprendizado de máquina (ML) voltados para a predição de Eventos Cardiovasculares Adversos Maiores (MACE). Este trabalho focou na padronização e organização dos dados para facilitar a análise e a melhoria dos modelos preditivos em um contexto multicêntrico e heterogêneo.

2.2. Objetivos Específicos

2.2.1 Revisão e Estudo de Métodos de ETL e Padronização de Dados:

Realizar uma revisão bibliográfica sobre processos de ETL (Extract, Transform, Load), o modelo OMOP-CDM (Observational Medical Outcomes Partnership - Common Data Model) e a importância da padronização de dados para estudos multicêntricos. Estudar procedimentos de ETL em pesquisas multicêntricas que utilizaram o OMOP-CDM para entender melhor as práticas e desafios envolvidos.

2.2.2. Implementação e Testes de ETL:

Apoiar a organização das coortes de dados dos hospitais vinculados ao HCFMRP e repositórios como o MIMIC-III. Implementar e testar procedimentos de ETL para o carregamento de dados em bancos de dados do HCFMRP e MIMIC-III, assegurando a correta integração e padronização das informações.

2.2.3. Modelagem e Padronização dos Dados:

Modelar os dados de acordo com o OMOP-CDM da OHDSI, incluindo a padronização da estrutura e vocabulário semântico para garantir consistência e interoperabilidade. Aplicar técnicas de ETL para garantir que os dados estejam prontos para análise em modelos de ML.

2.2.4. Integração e Estruturação de Dados:

Integrar dados anonimizados em um único repositório OMOP-CDM garantindo que todas as informações estejam consolidadas e acessíveis para análises futuras.

2.2.5. Aplicação de ETL para estruturar dados

Estudo de campos e aplicação de ETL em campos não estruturados visando a evolução dos modelos de AM voltados para a previsão de ocorrência de MACE.

3. MATERIAIS E MÉTODOS

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.

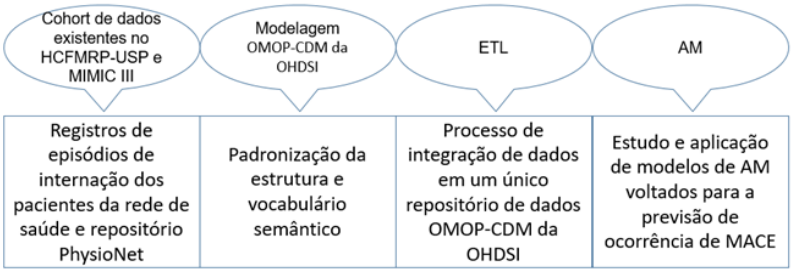

Figura 1: etapa dos materiais e métodos utilizados.

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. Com base no levantamento inicial, o HCFMRP-USP por meio do departamento de qualidade de dados disponibilizou um volume de 3.952,9 Gigabytes de dados clínicos. Os arquivos foram disponibilizados em formato “.csv”. Durante esta etapa, cada arquivo teve sua estrutura geral textual dividida em diferentes colunas em formato tabela, delimitando os caracteres separados de cada campo.

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. Foram implementados e testados procedimentos de ETL para o carregamento de dados nesses bancos, resultando na criação de um ambiente estruturado e padronizado conforme o modelo OMOP-CDM da OHDSI. A

padronização dos dados envolveu a modelagem da estrutura e do vocabulário semântico para garantir a interoperabilidade entre diferentes sistemas. Foram mapeados registros de pacientes e internações, com foco na criação de um banco de dados que facilitasse a integração e análise de dados clínicos e demográficos.

O registro identificador do paciente, por se tratar de um dado alfanumérico foi convertido para o formato texto, assim como todas as células de cada tabela com dados que envolvem combinações de números e letras; os dados que representam a indicação de tempo e período foram formatados para o tipo data, totalizando um conjunto de 14 arquivos. Operações básicas de ETL foram realizadas para anonimizar dados sensíveis; p.ex.: ao registro de identificação do paciente foi utilizado Anonymization & Data Masking for PostgreSQL[5].

Adicionalmente, foi utilizado o ambiente de integração de tecnologias Visual Code Microsoft, para possibilitar o uso da linguagem de programação Python com o banco de dados Postgres modelado segundo o OMOP-CDM da OHDSI.

Integração e Estruturação dos Dados

Sequentemente foi realizado o processo de integração dos dados anonimizados em um único repositório OMOP-CDM, utilizando Google BigQuery. Esse repositório consolidou informações provenientes de diferentes fontes e possibilitou a análise em larga escala. O carregamento dos dados no BigQuery permitiu a análise detalhada de informações demográficas dos pacientes, como idade, gênero e cidade de residência, que são essenciais para a modelagem de aprendizado de máquina. Foram realizadas diversas etapas de pré-processamento de dados para garantir a qualidade e a precisão das informações. Linhas duplicadas foram removidas, inconsistências foram corrigidas e os campos das planilhas foram padronizados para garantir uniformidade e integridade dos dados. Essas etapas foram fundamentais para preparar os dados para análise e inclusão em modelos preditivos.

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.

4. RESULTADOS E DISCUSSÕES

Com o uso das tecnologias foi possível implementar um ambiente estruturado e padronizado contendo o conjunto de dados do HCFMRP e do MIMIC-III. Foi realizada a criação bem-sucedida de um banco de dados padronizado conforme OMOP-CDM da OHDSI, com mapeamentos específicos de pacientes e internações (admissões), indicando bom potencial para garantir a interoperabilidade em pesquisas multicêntricas na área da saúde. A padronização viabilizou estudos sobre a generalização do desempenho de modelos de AM voltados para a previsão de ocorrência de MACE, considerando os dados das duas coortes.

O fluxo geral de ETL's está presente na Figura 2 e é detalhado na Figura 3. A ordem da execução das etapas é dada pela sequência definida pela representação esquemática de scripts por meio de dialetos para banco de dados relacional Postgres SQL.

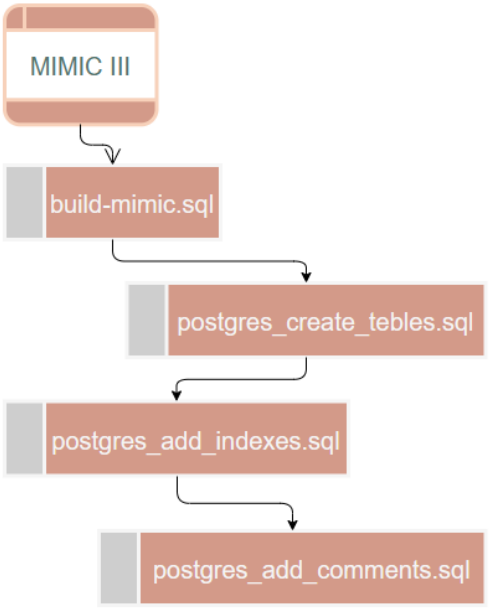

Figura 2: representação esquemática parcial do fluxo de ETLs.

A partir da execução parcial individual de cada script ETL, o ambiente de desenvolvimento de sistemas de informação organizado, integrado e automatizado para esta finalidade pode gerar um repositório de dados baseado no modelo OMOP-CDM. A Figura 4 apresenta a representação do ambiente de banco de dados MIMIC e a estrutura da tabela para armazenar dados de admissões de pacientes internados para investigação de ocorrência de MACE.

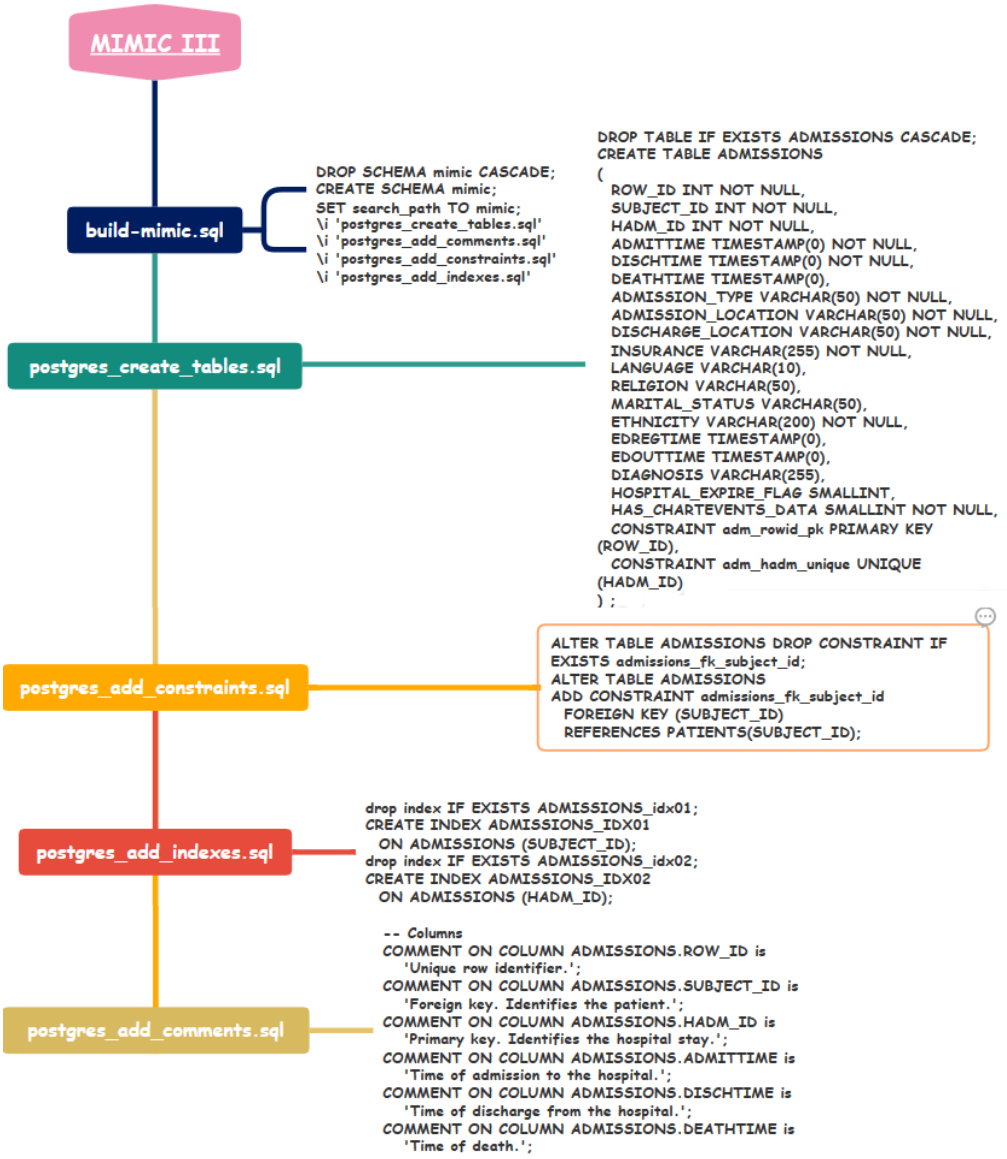

Figura 3: representação esquemática parcial do fluxo de ETLs.

O projeto contemplou um total de 27 tabelas (MIMIC-III) até o momento. Adicionalmente, foram elaborados ETLs complementares para integração de informações LOIC e um conjunto de tabelas em aderência com o modelo de normalização OMOP-CDM para geração de cohorts MACE. Até o presente momento da pesquisa, é possível atestar uma generalização de dados clínicos, anonimizados, no volume de 188,4 Megabytes. Trata-se de evidências em que o paciente pode aparecer mais de uma vez na base de dados por ter mais de uma internação no período. Durante esta etapa, foram estabelecidos dois grupos de pacientes: Grupo MACE e Grupo Controle.

Trata-se de evidências em que o paciente pode aparecer mais de uma vez nos arquivos por ter mais de uma internação no período de 01/01/2009 à 31/12/2022. Durante esta etapa, foram estabelecidos dois grupos de pacientes: Grupo MACE e Grupo controle.

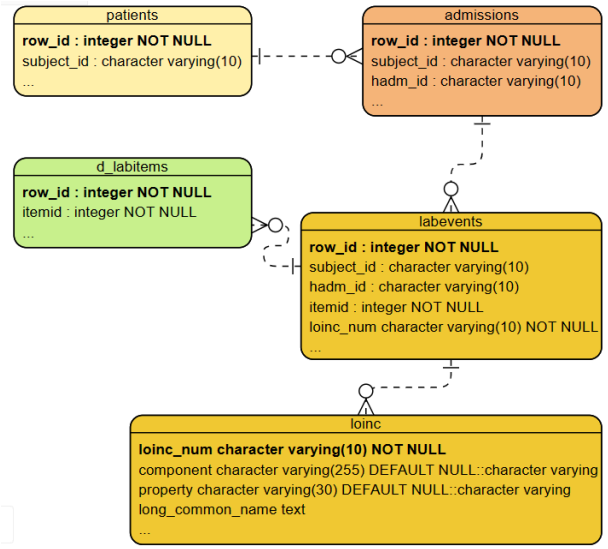

Figura 4: estrutura parcial do MIMIC III Postgres SQL.

A Figura 5, apresenta etapas realizadas durante o desenvolvimento utilizando a linguagem de programação Python conectado ao banco de dados MIMIC III e a figura 6 é o ambiente integrado utilizado no *Microsoft Visual Code*[6] (VsCode). A integração desta solução é possível com a importação de componentes de programação: VsCode.

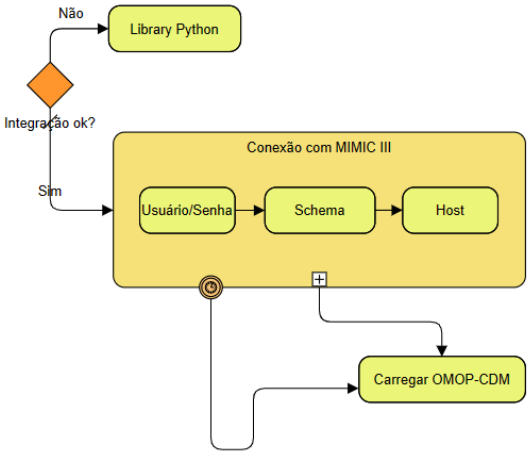

Figura 5: representação utilizada no ambiente de desenvolvimento *Microsoft Visual Code*.

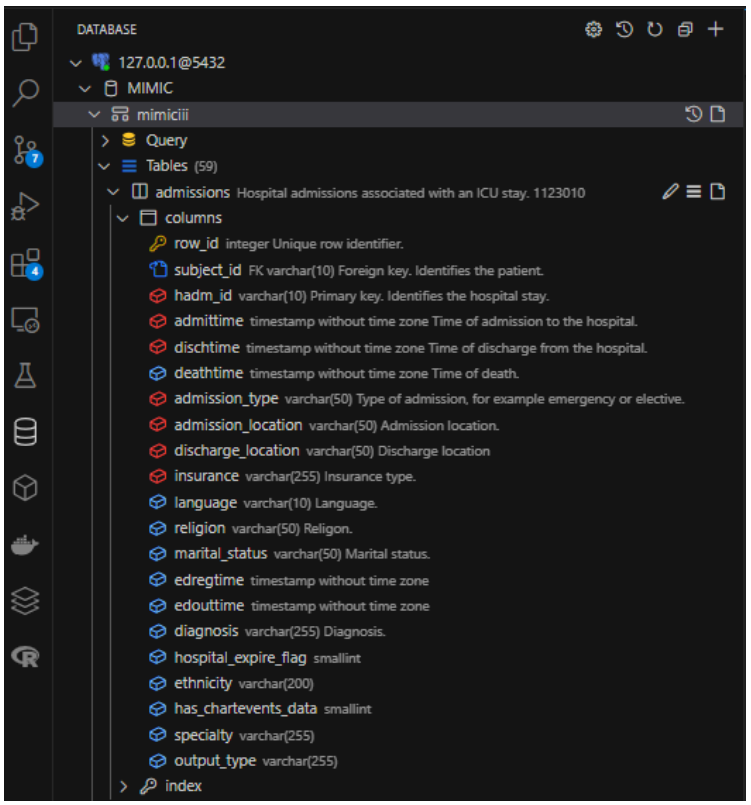

Figura 6: ambiente de desenvolvimento Python integrado com banco de dados MIMIC.

O treinamento e validação interna com os dados do HCFMRP resultou em uma área sob a curva ROC (AUC) de 0,87 para uma amostra de teste independente do próprio HCFMRP, para um classificador do tipo Random Forest. Contudo, ao aplicar o modelo treinado com os dados do HCFMRP a um conjunto de dados extraído da base do MIMIC-III, verificou-se uma redução na AUC para 0,70, indicando a necessidade de mais estudos voltados para a otimização da generalização do desempenho do modelo implementado para diferentes origens e populações. Esta discrepância destaca a importância de ajustes nos modelos para lidar com variações nos dados de diferentes bases. A diferença observada pode estar relacionada a diferenças nas características demográficas e clínicas entre as coortes, bem como à qualidade e à padronização dos dados. A interoperabilidade dos dados estruturados, embora garantida pelo uso do OMOP-CDM, requer ajustes adicionais nos modelos para manter a precisão preditiva em diferentes populações. Na figura 7 pode-se ver esses resultados graficamente.

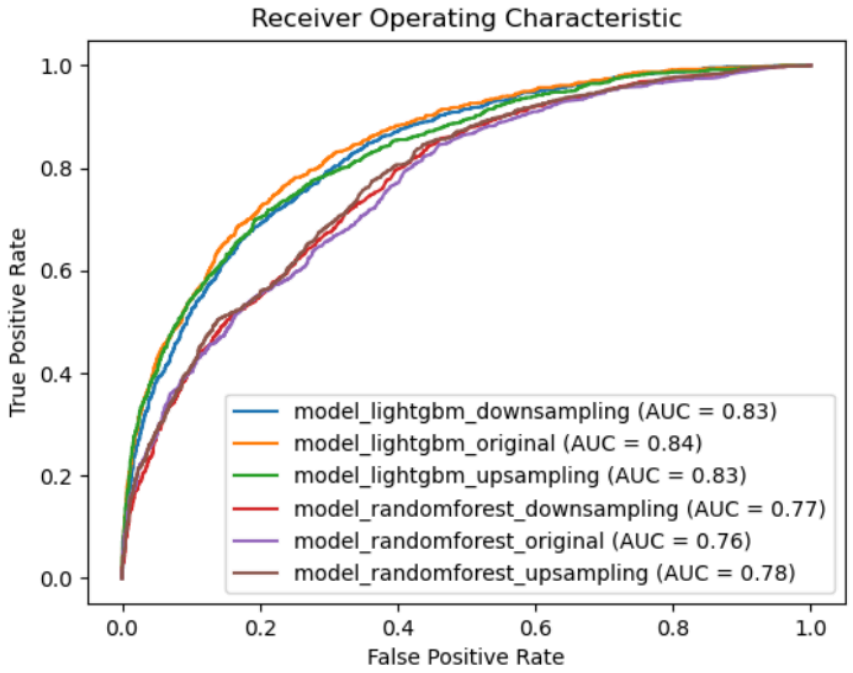

Figura 7: Desempenho do modelo sem utilizar os novos campos estruturados.

Concomitantemente, durante a estruturação do ambiente de generalização foi utilizado o algoritmo *Light Gradient Boosting (LGB)* para preparação e balanceamento de dados, assim como treinamento do modelo de AM. E neste momento, com 70% das amostras selecionadas aleatoriamente para treinamento dos modelos e 30% utilizadas para validação, o LGB apresentou um desempenho preditivo na validação interna (AUC = 0,871 (0,859 - 0,882); Precisão = 0,794 (0,782 - 0,808)) e externa (AUC =0,786 (0,778 - 0,792)).

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, contêm informações críticas, mas devido à falta de estrutura, são difíceis de categorizar de forma consistente para uso em modelos preditivos.

Os dados enviados pelo HCFMRP foram anonimizados utilizando uma técnica de pseudo-anonimização, permitindo que o processo pudesse ser revertido se necessário. Esses dados foram carregados no Google BigQuery, facilitando o armazenamento e a análise em larga escala. Através dos dados de internação (*admission*), foi possível extrair informações demográficas dos pacientes, como idade, data de óbito, gênero e cidade de residência. Esses dados demográficos são de extrema importância para os modelos de AM, pois permitem uma análise mais granular e precisa dos fatores que influenciam os eventos cardiovasculares. Além disso, utilizamos a tabela de diagnósticos, que contém os CIDs (Classificação Internacional de Doenças), para montar a coorte de MACE do HC, o que nos possibilitou identificar e categorizar com precisão os eventos cardiovasculares maiores.

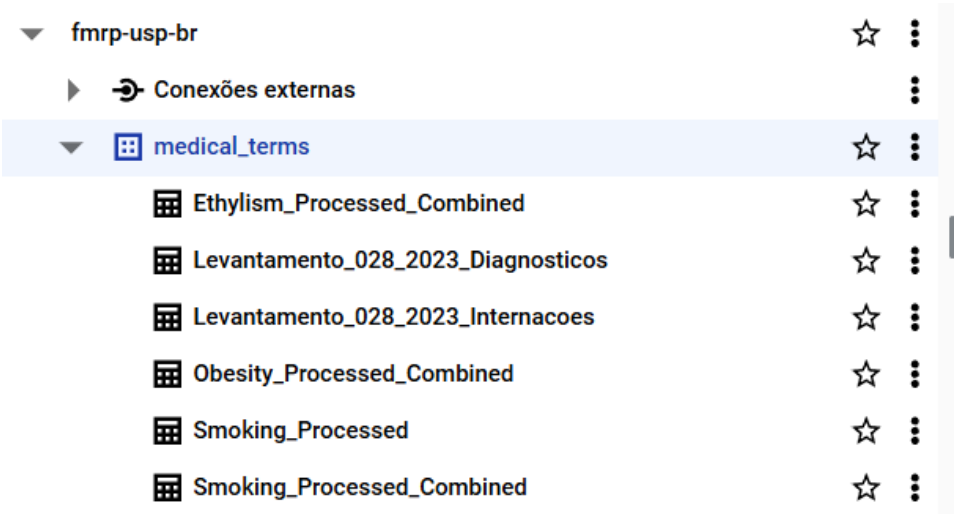

Figura 8: ambiente do Google BigQuery

Inicialmente, utilizamos um método de categorização simples baseado na presença de palavras-chave em campos de texto livre, como “tabagismo” e “tabagista”. Este método envolveu a aplicação de condições lógicas para atribuir rótulos como 'smoke' (fumante) e 'nosmoke' (não fumante), dependendo da ocorrência dessas palavras. Além disso, tentamos filtrar casos que continham negações explícitas, como "nega tabagismo" ou "negam tabagismo". No entanto, essa abordagem apresentou limitações significativas.

A principal limitação foi a incapacidade de capturar variações semânticas e negações menos explícitas, o que resultou em classificações incorretas ou inconsistentes. Por exemplo, frases como "cessou tabagismo" ou "ex-tabagista" indicam que o paciente é um ex-fumante, mas não estavam sendo adequadamente classificadas como 'nosmoke' na abordagem inicial. Isso evidenciou a necessidade de uma estratégia mais robusta para capturar a complexidade das expressões utilizadas nas notas clínicas. Dessa maneira, para abordar essas limitações, a estratégia de categorização foi aprimorada, criando uma nova coluna, chamada CATEGORY_2, na tabela Smoking_Processed_Combined, esta que já havia passado por um pré processamento inicial que removeu dados duplicados e nulos. Essa nova coluna foi gerada aplicando um conjunto de regras mais sofisticadas que incorporam uma gama ampla de expressões de negação e variações de linguagem.

O processo técnico envolveu o uso de SQL avançado no ambiente BigQuery, onde utilizamos a função STRING_AGG combinada com CASE WHEN para categorizar os textos com base em padrões mais complexos de detecção de palavras-chave e negações. Para garantir a correta interpretação das negações complexas, foi essencial capturar e categorizar expressões variadas que indicam um status de não tabagista ('nosmoke'), como "cessou tabagismo", "ex-tabagista", e "cessado tabagismo". Esses termos, embora diferentes, precisam ser reconhecidos como indicativos de cessação do tabagismo.

Além disso, incorporamos uma análise sensível aos contextos temporais, incluindo a interpretação de verbos e tempos verbais que refletem um histórico de tabagismo que não é mais relevante para o estado atual do paciente. Expressões como "negou tabagismo" e "cessado tabagismo" indicam uma condição passada que deve ser diferenciada de um status atual ativo de tabagismo. Foi também necessário tratar casos específicos onde as notas clínicas indicavam recomendações dos profissionais de saúde, que poderiam alterar a interpretação direta de cessação. Por exemplo, expressões como "cessação de tabagismo" precedidas por verbos como "oriento" sugerem que o paciente ainda não parou de fumar, mas recebeu orientação para fazê-lo.

Para evitar categorizações incorretas, as regras foram refinadas para não classificar erroneamente esses casos como 'nosmoke' com base apenas na presença do termo "cessação".

Finalmente, o processo incluiu atualizações contínuas para capturar de maneira mais abrangente as variações nos registros clínicos. Foram adicionadas condições para expressões como "%cessou tabagismo%", "%negou tabagismo%", "%cessado tabagismo%", "%negado etilismo ou tabagismo%", e "%passado de tabagismo%". Essas adições ampliaram a capacidade de reconhecimento e categorização das diversas formas de descrição do status de tabagismo dos pacientes, assegurando uma cobertura mais completa e precisa.

O código SQL usado para este refinamento pode ser lido na figura 9.

```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 9: Consulta utilizada para categorização.

A introdução das regras refinadas na coluna CATEGORY_2 proporcionou uma classificação mais robusta e precisa dos dados não estruturados sobre tabagismo. Esse avanço foi crucial para transformar dados textuais complexos em um target binário ('smoke' e 'nosmoke') utilizável por modelos de machine learning. A capacidade de categorizar corretamente essas informações possibilitou o uso dos dados não estruturados no treinamento dos modelos, contribuindo para a criação de predições mais precisas de MACE. Concomitantemente, está sendo exemplificado como técnicas de pré-processamento e regras de categorização refinadas podem impactar diretamente a eficácia dos modelos de machine learning na saúde, enfatizando a necessidade contínua de evoluir métodos de integração de dados para lidar com a complexidade dos dados clínicos não estruturados. Podemos perceber a melhora do modelo na figura 10 quando são incluídas novas informações .

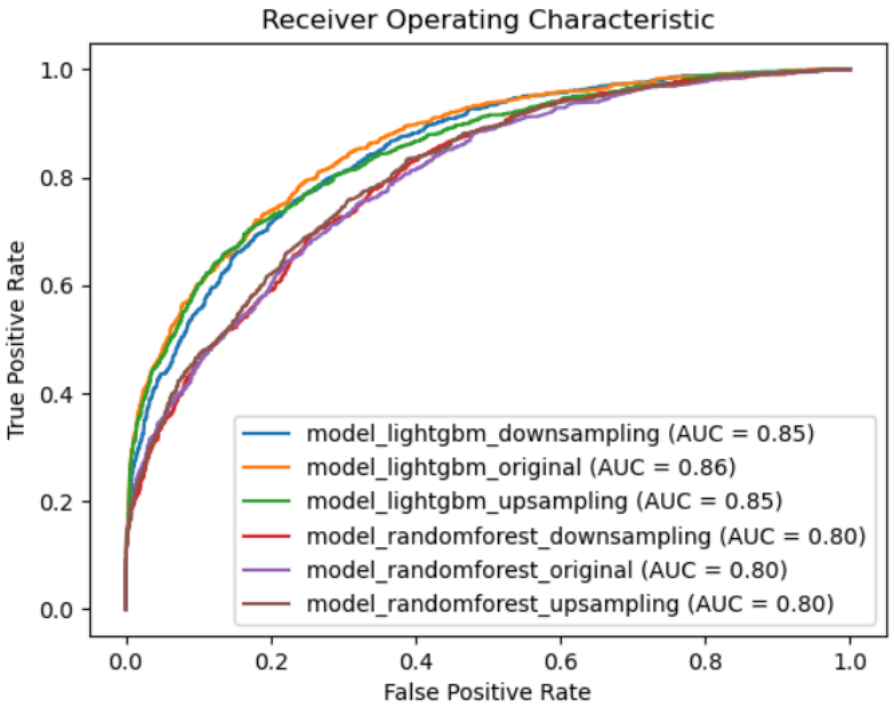

Figura 10: Resultados do modelo com dados demográficos, diagnósticos, laboratoriais, administrativos e médicos termos.

Não obstante, a importância da estruturação de dados é reafirmada pela necessidade de transformar informações não estruturadas em dados utilizáveis para modelos de machine learning. Esse processo não só enriquece os conjuntos de dados disponíveis para análise, mas também amplia a capacidade dos modelos de capturar nuances clínicas que podem ser essenciais para predições mais precisas e relevantes, especialmente em cenários multicêntricos onde a variabilidade entre populações é um desafio constante.

5. CONCLUSÃO

A implementação do OMOP-CDM em objetos específicos, como mapeamentos de pacientes e internações (admissões), representou um marco significativo neste projeto. A criação bem-sucedida de um banco de dados conforme OMOP-CDM da OHDSI, com mapeamentos específicos de pacientes e internações (admissões), ressalta a importância da padronização para garantir a interoperabilidade em pesquisas multicêntricas na área da saúde.

Foi de suma importância as categorizações de forma correta, afinal os dados classificados corretamente foram utilizados para treinar os modelos de AM. Um modelo de AM treinado com dados precisos e bem categorizados tende a ser mais robusto e eficaz, o que melhora a sua capacidade de previsão e a generalização para diferentes populações. Além disso, o aumento das variáveis de análise permite um estudo mais abrangente principalmente por estarmos falando de um evento adverso multifatorial.

Não obstante, a padronização dos dados conforme o OMOP-CDM facilitou a interoperabilidade entre diferentes instituições de saúde, permitindo a realização de estudos multicêntricos e a colaboração em larga escala. Isso não só fortalece a validação dos modelos de AM, mas também promove o avanço do conhecimento científico e a melhoria dos cuidados de saúde em nível global. O ganho de integrar duas bases de dados diferentes (HCFMRP e MIMIC-III) dentro dos padrões OMOP-CDM não pode ser subestimado. Ter acesso a essas duas bases distintas proporcionou uma oportunidade única de testar e validar os modelos de AM em dados diversos e heterogêneos. Essa diversidade de dados foi essencial para melhorar a capacidade dos modelos de AM de generalizar para diferentes populações e contextos clínicos.

Modelos que performam bem em múltiplas coortes são mais robustos e têm maior probabilidade de serem aplicáveis em cenários reais, o que é um passo crucial para a sua implementação em práticas clínicas. O impacto científico deste trabalho é substancial, pois demonstra como a padronização e a integração de dados não estruturados em registros eletrônicos de saúde podem melhorar a eficácia dos modelos preditivos de AM. Esse avanço não apenas contribui para a personalização dos cuidados médicos, mas também otimiza a utilização de recursos de saúde, potencialmente salvando vidas e reduzindo a morbidade associada a MACE.

A integração das bases do HCFMRP e do MIMIC-III no formato OMOP-CDM possibilitou a exposição dos modelos a uma diversidade de dados demográficos e clínicos, ampliando seu potencial de aplicação prática. Esse acesso a dados heterogêneos foi crucial para a validação dos modelos em populações distintas, enfrentando o desafio de generalizar predições para contextos clínicos variados. A interoperabilidade e padronização de dados são essenciais para evitar a perda de informações relevantes e garantir que os modelos de AM possam ser aplicados com precisão em diferentes cenários clínicos.

Uma das grandes dificuldades enfrentadas foi a integração de dados não estruturados, principalmente informações provenientes de notas clínicas sobre tabagismo, etilismo e obesidade. Esses registros, em campos de texto livre, apresentam uma grande variabilidade e complexidade semântica, como variações de tempo verbal e expressões de negação, por exemplo, "cessou tabagismo", "ex-tabagista", e "negou tabagismo". A categorização inicial, baseada em regras simples para identificação de palavras-chave, mostrou-se insuficiente, destacando a necessidade de abordagens mais refinadas.

A tabela Smoking_Processed_Combined exemplifica o avanço na integração de dados não estruturados, demonstrando que uma abordagem detalhada e criteriosa para a categorização

pode gerar alvos (targets) utilizáveis por modelos de machine learning. Esse processo possibilitou a criação de modelos de AM mais robustos e com maior capacidade de generalização, ao incluir variáveis que anteriormente estavam fora do escopo dos modelos tradicionais devido à natureza não estruturada dos dados.

Em conclusão, este projeto destaca a importância crítica da estruturação e categorização de dados não estruturados em registros eletrônicos de saúde para a construção de modelos preditivos eficazes na medicina. Os ganhos alcançados não só promovem o avanço da ciência médica, mas também apontam caminhos para futuras pesquisas na área de saúde pública e medicina de precisão, com o potencial de impacto global em sistemas de saúde diversificados.

7. REFERÊNCIAS

1. Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., & Nichols, M. (2016). Cardiovascular disease in Europe. 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. Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://physionet.org/content/mimiciii/1.4/
5. 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 />.

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

7. ((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.