**Método**

**Caracterização da pesquisa**

Esta pesquisa possui abordagem quantitativa, com uso de dados secundários e emprego de técnicas de *machine learning* para prever o desfecho do estudo: rotatividade (*churn*) de médicos. Neste sentido, foram acessados dados do 18º Ciclo do Programa Mais Médicos, realizado em 2019. A opção por este ciclo se deve pela presença de alguns atributos a nível indivíduo, que não estavam disponíveis em chamadas de ciclos anteriores. Já as chamadas de 2020 adiante não completaram um ciclo completo, considerando uma possível prorrogação. Todos os dados são públicos, disponíveis por meio de portarias do Ministério da Saúde, assim como bases de diversas fontes. Deste modo, não foi necessária a apreciação da investigação por comitê de ética em pesquisa. A seguir, narramos os passos usados para realizar o tratamento dos dados.

**Tratamento dos dados**

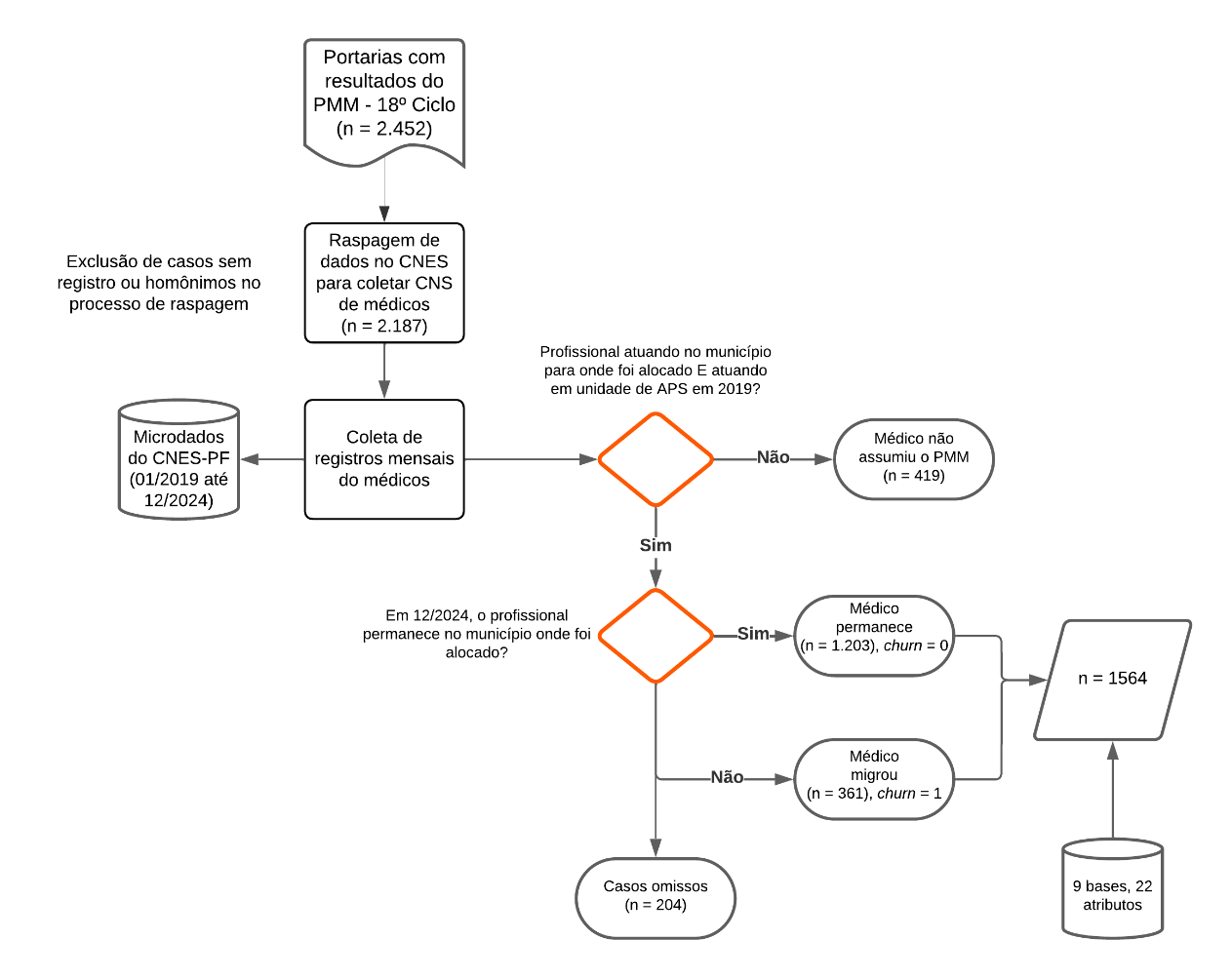
O primeiro passo consistiu em acessar dados dos médicos que foram aprovados do PMM no ciclo mencionado. Os resultados da seleção foram divulgados publicamente por meio da Portaria Nº 7, de 18 de junho de 2019, e a Portaria Nº 13 de 16 de agosto de 2019. Entre os dados divulgados e que foram utilizados para análise estão: nome, município para onde o profissional é alocado, data de nascimento, mês/ano de formação e participação pregressa no programa. Nesta etapa, o total de profissionais era de 2.452 médicos.

Em sequência, foi realizado um procedimento de raspagem de dados (webscrapping) na página do Cadastro Nacional de Estabelecimentos de Saúde-CNES (cnes.datasus.gov.br) com o objetivo de automatizar a coleta da identificação do profissional pelo número do Cartão Nacional de Saúde (CNS). Para este procedimento foi utilizada a biblioteca *RSelenium* na linguagem R. Um total de 256 médicos foram excluídos devido à falta de retorno nas consultas ou em razão da presença de homônimos nos retornos. Dessa forma, permaneceram 2.187 médicos para análise.

O próximo passo consistiu em acessar microdados da base CNES-Profissionais (CNES-PF) entre janeiro de 2019 e dezembro de 2024 que estão disponíveis via transferência de arquivos públicos do Datasus. O CNES-PF contém o registro mensal dos profissionais vinculados a estabelecimentos de saúde. Com isso, avaliamos alguns critérios para concluir a construção da base:

1. o profissional aprovado assumiu o programa? A resposta afirmativa dependia de duas condicionais: o profissional estar no município para onde o programa o designou em algum mês de 2019 e alocado em alguma unidade de Atenção Primária à Saúde (APS). O não atendimento a este conjunto de critérios o excluía da base, pois significava que, mesmo sendo aprovado, o indivíduo não assumiu a atribuição. Deste tratamento foram mantidos 1.768 médicos;
2. Passados cinco anos do programa, o indivíduo permanece no município ao qual foi alocado? Cabe lembrar que a duração do programa é de três anos mais, podendo ser prorrogado por mais um. A resposta para essa pergunta foi importante para definir a variável resposta do estudo. Se o indivíduo, em dezembro de 2024, permanece no mesmo município ao qual foi alocado, a variável assume o valor de “permanece”. Caso o profissional, em dezembro de 2024, esteja atuando em estabelecimentos de outros municípios, a variável assume o valor de “migrou”. Existem casos atípicos em que o profissional deixa de ter registros na base em anos/meses antes do término do programa. Esses casos foram excluídos, resultando uma amostra final de 1.564 médicos.

O tratamento completo dos dados pode ser sintetizado pela figura 1.



De posse da amostra, foi realizado o enriquecimento incluindo outros atributos em diferentes níveis de análise, conforme apresentado na fundamentação teórica. O quadro 1 lista as dimensões, variáveis e respectivas fontes de dados.

|  |  |  |
| --- | --- | --- |
| Dimensão | Variáveis | Fonte de dados |
| Individual | Município de alocação | Portaria Nº 7, de 18 de junho de 2019; Portaria Nº 13 de 16 de agosto de 2019 |
| Idade |
| Participação prévia no PMM |
| Estabelecimento | Quantidade média de agentes comunitários de saúde nas unidades de APS na qual o médico atuou | Cadastro Nacional de Estabelecimentos de Saúde (CNES) |
| Quantidade média de técnicos e auxiliares de enfermagem nas unidades de APS na qual o médico atuou |
| Quantidade média mensal de enfermeiros nas unidades de APS na qual o médico atuou |
| Quantidade média de cirurgiões-dentistas nas unidades de APS na qual o médico atuou |
| Vínculos de trabalho | Quantidade de vínculos com outros estabelecimentos de saúde (ex.: plantões em hospitais da região) |
| Tempo de atuação prévia, anterior ao PMM, no município alocado |
| Município | População | Instituto Brasileiro de Geografia e Estatística (IBGE) |
| PIB per capita de 2019 |
| Taxa de população ocupada |
| Número de vagas em curso de medicina | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) |
| Orçamento per capita em saúde | Índice de Desenvolvimento Sustentável das Cidades – Brasil (IDSC-BR) |
| Percentual da população atendida com esgoto sanitário |
| Taxa de homicídio |
| IDEB – Anos Iniciais |
| IDEB – Anos Finais |
| Investimento público em infraestrutura per capita |
| Número de equipamentos esportivos públicos por 100 mil habitantes |
| Número de centros culturais, espaços e casas de cultura públicos e privados por 100 mil habitantes |
| Distância da capital | Google Maps |

Fonte: elaborado pelos autores

**Análise de dados**

A modelagem preditiva com apoio de *machine learnin*g foi realizada utilizando linguagem Python utilizando o pacote sckit-learn. A variável de resposta foi o *churn* com valor um em caso de migração e zero em caso de permanência, tornando-se, dessa forma, um problema de classificação. Percorreu-se o processo de machine learning, no qual, uma das primeiras etapas foi pré-processamento dos dados. No caso de variáveis numéricas, realizou-se a substituição de valores faltantes pela mediana e a padronização pelo método Z-score. No caso das variáveis categóricas foi realizada a substituição dos valores ausentes pela categoria de maior frequência e a codificação utilizando o método de *one-hot encoding*. Os dados foram divididos em treino e teste, 80% e 20%, respectivamente.

Quatro modelos foram treinados e avaliados: Regressão Logística, Árvore de Decisão, *Random Forest*, XGBoost e LightGBM. A otimização dos hiperparâmetros foi realizada com GridSearchCV a partir de uma combinação definida previamente. O treinamento envolveu validação cruzada estratificada com StratifiedKFold em 5 partes. Devido ao desbalanceamento da variável resposta, utilizou-se a abordagem do *Synthetic Minority Oversampling Technique* (SMOTE). Ao final, utilizamos a biblioteca *Shapley Additive exPlanations* (SHAP) para adicionar uma camada de explicação das previsões do melhor modelo.