Filipe Penna Cerávolo Soares

Explorando Padrões de Mortalidade no Brasil com Aprendizado de Máquina

São Paulo, SP

2023

Filipe Penna Cerávolo Soares

Explorando Padrões de Mortalidade no Brasil com Aprendizado de Máquina

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro.

Universidade de São Paulo – USP

Escola Politécnica

Departamento de Engenharia de Computação e Sistemas Digitais (PCS)

Orientador: Prof. Dr. Artur Jordão

São Paulo, SP

2023

Catalogação-na-publicação

Soares, Filipe
Explorando Padrões de Mortalidade no Brasil com Aprendizado de Máquina / F. Soares -- São Paulo, 2023.
p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Mortalidade 2.Saúde Pública 3.Inteligência Artificial 4.Aprendizado Supervisionado I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Aos meus pais que me ensinaram a encarar desafios com um sorriso no rosto.

Agradecimentos

Gostaria de expressar minha gratidão a todos que contribuíram para a formação de um espírito curioso em mim e me nutriram com a confiança necessária para procurar as respostas das questões que florescem ao meu redor. Em primeiro lugar, agradeço meus pais e meu irmão por me proporcionarem um ambiente familiar propício para meu crescimento moral e intelectual. O exemplo presente de um sólido conjunto de valores morais e vasto conhecimento acumulado foi uma grande inspiração para meu percurso.

Além disso, sou imensamente grato ao meu professor orientador, Prof. Dr. Artur Jordão, que guiou meu caminho acadêmico e me ajudou a expandir meu conhecimento nesse estudo. Também agradeço à Escola Politécnica da USP e à École Centrale Lyon, onde enfrentei meus maiores desafios acadêmicos, em ambientes enriquecedores do ponto de vista intelectual e humano.

No fim de uma longa jornada, saboreio o privilégio de ter sido apoiado por muitos, por ter muito aprendido e aguardo os desafios do futuro, ansioso a descobrir o que me revelarão.

Resumo

Este trabalho apresenta um estudo de aplicação de aprendizado de máquina, em bases de dados públicas disponíveis para toda a população, em casos de óbitos no Brasil. Foram desenvolvidos modelos de aprendizado supervisionado de *boosting* e arquiteturas modernas de redes neurais que correspondem ao estado atual da arte. Com base em suas implementações, observou-se a possibilidade de prever as causas de óbito. Além disso, foi reconhecido que esses modelos podem generalizar as previsões feitas para amostras externas, para os conjuntos de treinamento e teste, que foram constituídos por apenas um estado, para outros estados brasileiros. Finalmente, por meio de uma análise de seleção de atributos, foi possível determinar atributos que têm mais influência na correta previsão das categorias de mortalidade e uma porcentagem que apresenta uma melhor relação entre custo computacional e qualidade da previsão.

Palavras-chave: Mortalidade, Saúde Pública, Inteligência Artificial, Aprendizado Supervisionado.

Abstract

This work presents a machine learning application study, on public databases available to all population, to decease cases in Brazil. Supervised learning models of boosting and modern neural networks archtectures that correspond to current state of art were developed. Based on their implementation, it was seen that it was possible to predict decease causes. Besides, it was also recognized that these models can generalize the predictions made to external samples, to the train and test datasets, which were constituted by only one state, to other brazilian states. Finally, through a features selecion analysis, it was possible determine features that possesses more influence on the correct prediction of mortality categories and a percentage that presents a better relation of computation cost and prediction quality.

Keywords: Mortality, Public Health, Artificial Intelligence, Supervised Learning

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Motivação

Nas últimas décadas, foi evidenciado como as causas de morte nos Estados Unidos refletem uma complexa interação entre fatores sociais, econômicos, biológicos e comportamentais (CHANG et al., 2016). Esse contexto é reflexo de uma nação multiétnica e multicultural, virtude de um processo de povoamento marcado por imigrações e choques culturais. Desde então, compreender essas causas e as disparidades que existem entre diferentes grupos de pessoas tem sido fundamental no país para informar políticas de saúde pública e iniciativas destinadas a melhorar a saúde e o bem-estar da população.

De forma análoga, uma sociedade igualmente complexa formou-se no Brasil, por meio de processos históricos que guardam, até certo ponto, similaridade com os Estados Unidos. Em virtude desses fatores, pesquisadores investigaram a correlação de alguns fatores com causas de mortalidade, como a cor da pele (BATISTA; ESCUDER; PEREIRA, 2004). Isso sugere que os dados possam revelar indícios das causas das mortes, com bases em elementos do indivíduo, do seu contexto socioeconômico, e das causas do óbito. Por outro lado, o processo de classificação das mortes é realizado por pessoas e laudos manuais, auxiliados ou não por autópsia, razão pela qual está suscetível a erros. O desenvolvimento de um sistema de rotulação automática da causa da morte do indivíduo que auxilie os profissionais da saúde no processo de registro das causas das mortes poderia reduzir a incidência de erros ou permitir uma possível revisão de casos, caso convenha.

1.2 Objetivos

O objetivo desse estudo corresponde a automaticamente predizer as principais causas de mortalidade no Brasil sob a ótica de modelos de aprendizado de máquina supervisionado, a fim de compreender as possibilidades e limites dessas técnicas nesse contexto.

1.2.1 Questões de Pesquisa

Em particular, ao longo do desenvolvimento deste trabalho, as seguintes questões de pesquisa foram exploradas:

(i) **Seria possível predizer a causa de morte de um indivíduo, a partir de um conjunto de dados que reflita particularidades do paciente e do óbito**

registrado? Dado que o processo de rotulação de dados é um processo árduo e suscetível a erros, responder esse questionamento positivamente possibilita benefícios e avanços rumo a uma rotulação mais confiável e ágil. A fim de discutir essa questão de pesquisa, esse projeto utiliza técnicas preditivas supervisionadas de alta performance (estado da arte) baseadas em *boosting*, como *XGBoost, CatBoost*, e *LightGBM* e baseadas em redes neurais profundas, como *Transformer* e MLP Residual *(Residual Multi-Layer Perceptrons)*.

(ii) **Os modelos desenvolvidos generalizam as predições obtidas para amostras externas aos casos de treinamento e teste?** Durante o projeto com esses modelos de alta performance, foi necessário realizar recortes no espaço amostral para viabilizar o estudo em tempo hábil e com recursos computacionais limitados. Em particular, foi necessário realizar o estudo com dados de apenas um estado. Portanto, é importante entender se o modelo treinado para esse estado mantém sua capacidade preditiva para outros estados.

(iii) **Existem atributos que desempenham maior impacto na predição da causa da morte?** Seria, por exemplo, a idade o fator mais relevante para uma determinada causa de morte, ou seria a ocupação do individuo? Ou ainda, o dia da semana em que o indivíduo nasceu. Essa direção de pesquisa pode auxiliar autoridades da saúde e outros órgãos a dar maior atenção a determinados apontadores de qualidade de vida. Para este propósito, foram utilizados os métodos *Inf-FS Supervised e Inf-FS Unsupervised* (ROFFO SIMONE MELZI, 2020) que estão entre as técnicas do estado da arte em seleção de atributos.

1.3 Justificativa

Conforme exposto anteriormente, o estudo da mortalidade é um recurso importante para a criação e orientação de políticas públicas.

O uso de técnicas de aprendizado de máquina permite uma nova abordagem para essa questão, já que proporciona uma perspectiva complementar e holística sobre as causas de mortalidade e permite identificar fatores que podem não ser evidentes isto é, latentes em análises tradicionais. Além disso, o modelo supervisionado pode auxiliar em processos suscetíveis a erros, como a rotulação das causas de morte, na etapa de tratamento de óbito.

1.4 Organização do Trabalho

No Capítulo 2 são apresentados os aspectos conceituais e a organização do estudo. Nos aspectos conceituais, inicia-se pela terminologia adotada nesse trabalho, em seguida é

apresentado formalmente os principais conceitos estudados, como o aprendizado supervisionado e cada uma das técnicas utilizadas. Além disso, também é discutida a calibragem de hiperparâmetros dos modelos preditivos, seleção de atributos e avaliação de performance. Na parte dedicada à organização do estudo, apresenta-se as fases do projeto, bem como as atividades envolvidas nessas fases.

O Capítulo 3 discute todo o desenvolvimento das questão de pesquisa. Primeiro apresenta-se as configurações experimentais nas quais o projeto foi desenvolvido. Em seguida, discute-se o projeto e a implementação, passando pela fonte e extração de dados, seguida pelo tratamento desses dados necessários para conduzir corretamente os experimentos. Finalmente, nessa seção todos os experimentos são realizados e as questões de pesquisa são postas à confirmação empírica.

O Capítulo 4 relate as conclusões do projeto de formatura, bem como as perspectivas de continuidade do projeto.

1.5 Contribuições

Nesse trabalho foi desenvolvido um modelo preditivo da causa de morte de novas amostras que pode ser diretamente aplicado para tomada de decisões tanto no processo de classificação de causas de óbito quanto na tomada de decisão de políticas públicas.

Além disso, o estudo desenvolvido para seleção de atributos, traz mais informações sobre a importância relativa de cada, o que pode enriquecer o conhecimento dos profissionais da área sobre a questão.

1.6 Trabalhos relacionados

Em (LEMES; LEMOS, 2020), apresenta-se um mapeamento das iniciativas de inteligência artificial na esfera da saúde pública. Em (RIBEIRO et al., 2021), discute-se como a inteligência artificial impacta a rotina dos profissionais da saúde e como eles podem utilizá-la para complementar sua atuação. Em (SOUZA; BULGARELI, 2023), apresenta-se o uso de inteligência artificial aplicada ao processo decisório na alocação de recursos na saúde pública do Brasil, com discussão de casos de uso.

2 Metodologia

2.1 Aspectos Conceituais

O aprendizado de máquina (*machine learning*) é um conjunto de diferentes técnicas para realizar tarefas cognitivas como, por exemplo, processamento de linguagem natural.

Tais técnicas são baseadas na ideia de que máquinas podem aprender a tomar decisões, por meio da experiência, assim como humanos. Formalmente, um programa computacional é dito que aprende pela experiência E com respeito a um conjunto de tarefas T e desempenho P, se o seu desempenho nas tarefas em T, medido por P, melhora com a experiência E (MITCHELL, 1997).

Experiência no universo digital existe na forma de dados. Por isso, de maneira mais objetiva, o aprendizado de máquina pode ser descrito como o processo de (i) reunir um conjunto de dados (*data set*) e (ii) construir um modelo preditivo baseado nesse conjunto de dados, de modo a se resolver um dado problema (BURKOV, 2019).

2.1.1 Terminologia

O conjunto de dados é composto por *instâncias* ou *amostras* que descrevem o objeto ou evento de análise em diferentes dimensões. Cada uma dessas dimensões é conhecido como *atributo* (*attributes*) ou *atributos* e os seus respectivos valores numa dada amostra são conhecidos como *valor da amostra*.

Formalmente, seja $D = x_1, x_2, ..., x_m$ um data set com m instâncias, em que cada instância é descrita por d atributos. Logo, $\forall i \in [1; m],\ x_i = (x_i^1; x_i^2; ...; x_i^d)$, em que $\forall i \in [1; m],\ \forall j \in [1; d],\ x_i^j$ é o valor da amostra i no atributo j. Alternativamente, $D = x^1, x^2, ..., x^p$, em que x^p é o *vetor de atributos*, tal que $\forall j \in [1; d],\ x^j = x_1^j, x_2^j, ...x_m^j$.

2.1.2 Aprendizado supervisionado

O aprendizado supervisionado corresponde a uma das principais categorias de modelos de aprendizado de máquina. Nele, o conjunto de dados é uma coleção de exemplos rotulados, ou seja, $D = (x_i, y_i)_{i=1}^m$, em que $\forall i \in [1; m], y_i$ corresponde ao rótulo da amostra i. Esse rótulo pode ser um valor numérico (ou qualquer estrutura numérica) ou uma categoria. Neste trabalho, o rótulo/categoria corresponde a causa da morte de um indivíduo atribuído por um processo manual.

Sejam A e B dois domínios, um problema de predição consiste em estabelecer um mapeamento entre valores em um domínio, para valores em outro domínio por meio de

uma função: $f : A \rightarrow B$, na qual $f(a) = b$, onde a e b são, respectivamente, amostras do domínio A e B.

Em modelos de aprendizado de máquina, um conjunto de amostras é selecionado a fim de estimar a função f. Esse conjunto de amostras são conhecidos respectivamente como conjunto de treinamento. Em seguida, essa função é aplicada a um novo conjunto de amostras a fim de estimar os valores no domínio de destino (b).

Seja n, tal que $n < m$, a etapa de treinamento consiste em selecionar um subconjunto de amostras de D de tal forma que $D_{treinamento} = (x_i, y_i)_{i=1}^{n}$, com $X_{treinamento} = (x_i)_{i=1}^{n}$ e $y_{treinamento} = (y_i)_{i=1}^{n}$, a fim de modelar uma função $g : X \rightarrow y$ capaz de estipular uma relação entre os vetores de atributos com o rótulo correspondente. Com g definida, é possível aplicá-la sobre os vetores de atributos de $D_{teste} = (x_i, y_i)_{i=n+1}^{m}$, especificamente em $x_{teste} = (x_i)_{i=n+1}^{m}$ para estimar os rótulos $y_{teste} = (y_i)_{i=n+1}^{m}$.

O cálculo da função g, na etapa de treinamento e consequentemente sua aplicação na etapa de testes, pode ser realizada de várias formar diferentes e depende da técnica supervisionada de aprendizado escolhida.

No caso do estudo conduzido neste trabalho, conceitualmente os atributos relevantes podem tratar sobre duas categorias de informações diferentes: (i) ao indivíduo que veio a óbito ou (ii) a crise que o levou a óbito. Sobre o indivíduo, podem ser relevantes dados de diversas naturezas, como educação, gestação pela qual passou, contexto familiar e outros indicadores socioeconômicos. Já sobre a crise, é relevante entender se o paciente foi submetido a exames antes de falecer, se passou por procedimentos cirúrgicos ou por uma autópsia, depois que falecido.

Já os rótulos a serem preditos são as causas de morte, que conceitualmente serão categorias que refletem condições de saúde do indivíduo. Assim:

$$h : X_{falecido} \cup X_{crise} \rightarrow y_{causa\ morte}.$$

Sendo assim, o trabalho realizado no estudo consistiu em separar um subconjunto de dados para treinamento, estimar sobre ele uma função h que traduzisse a relação entre informações de um falecido, da crise que o levou a óbito e a causa de sua morte. Em seguida, aplicar essa função sobre uma parte ou todo o subconjunto de dados restante e estimar a causa da morte das amostras, em outras palavras, do indivíduo.

2.1.2.1 Técnicas de aprendizado supervisionado

Dois dos conjuntos de técnicas de aprendizado supervisionado mais relevantes atualmente são aquelas baseadas na técnica de boosting e em redes profundas (HOLLMANN SAMUEL MüLLER, 2023).

Modelos baseados em boosting

Os modelos de boosting são baseados na ideia de que modelos de aprendizado básicos ou fracos, ou seja, modelos simples que aprendem pouco individualmente *weak learners* e portanto tem uma capacidade preditiva limitada. No entanto, quando combinados, possuem elevada capacidade preditiva e levam a resultados positivos. O benefício dessa técnica é que essa combinação maximizam os resultados, como em modelos mais complexos individualmente, mas com menos *bias* e *variância*. Ou seja, com uma capacidade melhor de generalização da performance para amostras externas ao subconjunto de dados de treino e teste.

Os exemplos mais modernos e performantes desse tipo de modelos são baseados no *gradient boosting*. O algoritmo funciona construindo uma série de árvores de decisão, cada uma das quais é treinada para reduzir o erro da previsão anterior. O algoritmo possui essência iterativa, começando com uma previsão inicial. Em seguida, uma árvore de decisão é treinada para reduzir o erro da previsão anterior. A previsão atual é então atualizada para incluir a contribuição da nova árvore de decisão. Este processo é repetido até que o erro de previsão seja aceitável ou até que um número máximo de árvores tenha sido construído.

Sendo Z o número de passos definido para o algoritmo, $\forall z \in [1, Z]$, ele pode ser expresso formalmente da seguinte forma (LI,):

$$F_{z+1}(x) = F_z(x) + h_z(x) = y,$$

ou alternativamente:

$$h_z(x) = y - F_z(x).$$

Onde:

- $F(x)$ é uma função que estima de forma imperfeita a relação entre os domínios
- $h_z(x)$ é um estimador que será ajustado ao resíduo $y - F(x)$

O algoritmo é relativamente fácil de implementar e é eficiente em termos de tempo de treinamento. Além disso, é um algoritmo poderoso que pode ser usado para resolver uma variedade de problemas de aprendizado de máquina.

De maneira mais formal, o algoritmo busca reduzir uma função de custos em um problema de otimização. Em outras palavras, o objetivo é encontrar uma aproximação $\hat{F}$ para uma função $F(x)$ que minimiza o valor esperado de alguma função de perda especificada $L(y, F(x))$:

$$\hat{F} = \arg\min_{h} \mathbb{E}_{(x,y)}[L(y, F(x))].$$

Isso é feito por meio de uma soma ponderada de funções $h_i(x)$, que pertencem a uma determinada classe de modelos de aprendizagem simples.

$$\hat{F} = \sum_{i=1}^{Z} \gamma_i h_i(x) + const.$$

Em que:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_z(x) = F_{z-1}(x) \arg\min_{h_z} \left[\sum_{i=1}^{n} L(y_i, F_{z-1}(x_i) + h_m(x_i) \right]$$

Escolher a melhor função h em cada etapa para uma função de perda arbitrária L é, em geral, um problema de otimização computacionalmente inviável. Diante dessa complexidade, a abordagem é restrita a uma versão simplificada do problema. A estratégia consiste em aplicar uma etapa de máximo declive a esse problema de minimização, conhecida como descida de gradiente funcional. No caso contínuo, quando $\mathcal{H}$ é o conjunto de funções diferenciáveis arbitrárias em $\mathbb{R}$, obtém-se a seguinte equação:

$$F_z(x) = F_{z-1}(x) - \gamma_m \sum_{i=1}^{n} \Delta_{Fz-1} L(y_i, F_{z-1}(x_i)$$

em que as derivadas são obtidas com relação às funções F_i para $i \in \{1, \ldots, m\}$ e γ_m é o comprimento do passo. No entanto, no caso discreto, ou seja, quando $\mathcal{H}$ é finito, escolhe-se a função candidata h mais próxima do gradiente de L para a qual o coeficiente γ pode então ser calculado com o auxílio da pesquisa linear nas equações acima. Essa abordagem é uma heurística e, portanto, não produz uma solução exata para o problema em questão, e sim uma aproximação.

Apesar de não ser restringido, geralmente o *gradient boost* é aplicado em árvores de decisão, ou seja as funções F calculadas, são feitas nesse formato de algoritmo. Árvores de decisão criam um modelo que prevê o rótulo ao avaliar uma árvore de perguntas de características verdadeiro/falso do tipo se-então-senão, e estimam o número mínimo de perguntas necessárias para avaliar a probabilidade de tomar uma decisão correta. Árvores de decisão podem ser usadas para classificação, prevendo uma categoria, ou para regressão, prevendo um valor numérico contínuo.

XGBoost

Acrônimo para *Extreme Gradient Boost* é um método eficiente e amplamente utilizado atualmente, conquistou muito destaque pela performance em competições de

aprendizado supervisionado, se tornando a principal biblioteca para regressão, classificação e problemas de ranqueamento (NVIDIA,). É uma implementação do *Gradient-Boosted Decision Tree* (GBDT) (gradient boost aplicável em árvores de decisão).

LightGBM

Acrônimo para *Light Gradient-Boosting machine*, é outra implementação de GBDT.

LightGBM compartilha muitas das vantagens do XGBoost, como otimização esparsa, treinamento paralelo, múltiplas funções de perda, regularização, *bagging* e parada antecipada. Uma diferença significativa entre os dois está na construção das árvores. Ao contrário da abordagem de crescimento nível a nível — linha por linha — como a maioria das outras implementações, o LightGBM cresce árvores de forma orientada às folhas. Ele escolhe a folha que acredita proporcionar o maior decréscimo na perda. Além disso, o LightGBM não utiliza o algoritmo amplamente usado de aprendizado de árvores de decisão baseado em valores de características ordenados, que busca o melhor ponto de divisão em valores de características ordenados, como faz o XGBoost ou outras implementações. Em vez disso, o LightGBM implementa um algoritmo de aprendizado de árvores de decisão baseado em histograma altamente otimizado, proporcionando grandes vantagens em eficiência e consumo de memória. O algoritmo LightGBM utiliza duas técnicas inovadoras chamadas Amostragem Unilateral com Base no Gradiente (GOSS) e Agrupamento Exclusivo de Características (EFB), que permitem que o algoritmo seja mais rápido, mantendo um alto nível de precisão.

CatBoost

Acrônimo para *Categorical Boost* (boost categórico). Corresponde a outra implementação de GBDT.

Usa árvores de decisão binárias como preditores base. Durante o treinamento, ele cria um conjunto de árvores de decisão continuamente. Cada árvore seguinte é construída reduzindo a perda comparado com com as anteriores. Uma das principais diferenças entre o CatBoost e os demais algoritmos é sua implementação de *symmetric trees* (ou *oblivious tree*). O termo Oblivious significa que o mesmo critério de divisão é usado em todo o nível da árvore. Essas árvores são balanceadas, sendo menos propensas a *overfitting* e permitem acelerar significativamente a execução do modelo no momento do teste.

Outras duas características que diferem em relação aos outros dois tipos são o boosting ordenado dos dados e as permutações aleatórias dos dados.

Redes neurais profundas

Redes neurais também são uma alternativa popular e com alta performance para problemas de aprendizado supervisionado atuais.

As redes neurais são fundamentadas na simulação do funcionamento do cérebro humano, compreendendo uma rede interconectada de neurônios artificiais. Cada neurônio, ou unidade, é comparável a um *weak learner*, sendo um componente simples com capacidade preditiva limitada. No entanto, ao conectar essas unidades em camadas e treiná-las em conjunto, as redes neurais adquirem uma capacidade preditiva elevada, resultando em desempenho positivo.

A principal vantagem dessa abordagem reside na capacidade de modelar relações complexas nos dados, assim como em modelos mais robustos, mas com menor viés e variância. Isso implica uma melhor generalização do desempenho para conjuntos de dados externos ao treinamento e teste originais.

Atualmente, as arquiteturas mais avançadas e eficazes de redes neurais baseiam-se em técnicas como *deep learning*. O treinamento ocorre através da propagação do erro pela rede, ajustando os pesos das conexões entre neurônios para minimizar as discrepâncias entre as previsões e os resultados reais. A natureza iterativa desse processo permite que a rede aprenda padrões complexos e adapte-se a nuances nos dados.

Transformer

O Transformer é um modelo de *deep learning* de estado da arte baseado no mecanismo de atenção (VASWANI et al., 2017). Quando proposto, o modelo foi uma alternativa mais paralelizável e mais rápida em relação a outros modelos de redes neurais profundas com convolução e baseados em *boosting* para certos benchmarks.

Isso acontece em virtude da arquitetura de atenção que emprega que permite, ao contrário de abordagens convencionais que dependem fortemente de operações sequenciais, que permite a contextualização eficiente de cada elemento em uma sequência em relação a todos os outros. Isso não apenas melhora a capacidade de capturar relações de longo alcance, mas também possibilita treinamentos mais rápidos e eficientes.

MLP Residual

Como mencionado anteriormente, as redes neurais profundas têm demonstrado sucesso ao lidar com tarefas complexas de aprendizado supervisionado. No entanto, à medida que a profundidade da rede aumenta, surgem desafios relacionados à otimização do treinamento e à mitigação do desaparecimento do gradiente. O MLP Residual (TOUVRON et al., 2023) aborda essas questões de uma maneira única, introduzindo conexões residuais que permitem o fluxo direto de informações através das camadas.

A arquitetura residual do MLP permite que a rede aprenda não apenas as características discriminativas dos dados, mas também as diferenças residuais, facilitando o treinamento de redes mais profundas sem comprometer a eficácia. Essa estrutura inovadora é especialmente valiosa para lidar com conjuntos de dados complexos nos quais as relações entre as características são intrincadas.

Ao incorporar módulos residuais em cada camada, o MLP Residual atenua o desafio do desvanecimento do gradiente, permitindo uma propagação mais eficiente dos gradientes durante o treinamento. Isso resulta em uma convergência mais rápida e em modelos mais estáveis, contribuindo para o avanço do desempenho em uma variedade de tarefas.

2.1.2.2 Calibragem de hiperparâmetros

Existem nos modelos de aprendizado de máquina, parâmetros que são definidos antes da etapa de testes e não são atualizados durante esse processo. Eles são conhecidos como *hiperparâmetros* e podem ser calibrados a fim de maximizar a performance do modelo. O processo que realiza isso é conhecido como *otimização de hiperparâmetros*, ou simplesmente calibragem de hiperparâmetros.

Para testar os hiperparâmetros, existem estratégias exaustivas como o *grid search* ou *parameter sweep*, que testem todos os valores em faixas de dados. No entanto, esses algoritmos são excessivamente custosos computacionalmente. Por isso, foram desenvolvidas soluções mais eficientes para encontrar os melhores valores na faixa de exploração.

A otimização Bayesiana é um método de otimização global que constrói um modelo probabilístico da função que mapeia os valores dos hiperparâmetros para o objetivo avaliado em um conjunto de validação. Ao avaliar iterativamente uma configuração promissora de hiperparâmetros com base no modelo atual e, em seguida, atualizá-lo, a otimização bayesiana visa reunir observações que revelem o máximo de informações possível sobre essa função e, em particular, a localização do ótimo.

2.1.2.3 Seleção de atributos

A qualidade da predição do modelo reflete a qualidade dos dados nos vetores de atributos. A inclusão de um atributo adicional em um modelo vai necessariamente aumentar o custo computacional de treino e teste do modelo sem necessariamente implicar um ganho proporcional. Além disso, não importa em qual modelo de predição adotado, existem parâmetros que são mais ou menos importante para a correta predição dos valores.

A fim de metrificar a relevância de cada parâmetro em um modelo e assim poder tomar ações no sentido de otimizar os resultados, foram desenvolvidos algoritmos de seleção de atributos. Esses algoritmos analisam a contribuição de cada atributo na capacidade preditiva do modelo, identificando aqueles que têm um impacto significativo e descartando os menos relevantes. Dessa forma, a seleção de atributos não apenas melhora a eficiência computacional, reduzindo o custo de treino e teste, mas também aprimora a generalização do modelo, evitando overfitting e promovendo um desempenho mais robusto em dados não vistos.

Neste projeto, são utilizados os algoritmos de estado da *Infinite Attribute Selection*

(IFS) supervisionados e não supervisionados (Inf-FS$_S$ e Inf-FS$_U$ respectivamente) (ROFFO SIMONE MELZI, 2020). Eles são baseados em uma estratégia de filtro dos atributos por meio de uma abordagem de grafos que ranqueia e seleciona os atributos com base nos possíveis conjuntos de atributos como caminhos no grafo.

2.1.2.4 Avaliação de performance

Uma das métricas mais relevantes para avaliação de um modelo é a acurácia. Ou seja, qual a porcentagem dos valores corretamente classificados em suas categorias. Seja $\hat{y}_i$ a predição da i-ésima amostra e y_i seu valor real, a fração de predições corretas (acurácia) sobre n amostras é definida da seguinte forma:

$$acurácia(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (\hat{y}_i = y_i)$$

Quando existem muitas categorias, como no presente caso de estudo, existe uma métrica conhecida como *acurácia top-k*. Ela é uma generalização da acurácia, mas ela não se restringe a avaliar a fração das amostras em que a classe com a maior probabilidade predita é a correta, mas em que a classe correta está dentro das k categorias mais prováveis. Sendo $\hat{f}_{i,j}$ a classe predita para a i-ésima amostra correspondente a j-ésima maior classe mais provável e y_i a classe real da amostra, a acurácia top-k sobre n amostras é definida por:

$$acurácia\ top-k(y, \hat{f}) = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{i=0}^{n-1} (\hat{f}_{i,j} = y_i)$$

3 Desenvolvimento do Trabalho

3.1 Configurações experimentais

Para o desenvolvimento do projeto, foi utilizado Python em diferentes versões e sistemas operacionais em uma máquina equipada com um processador Intel i7-12700H (12a geração), que possui 12 núcleos de processamento.

Todo o projeto foi desenvolvido utilizando o GitHub em modo público e a integralidade dos códigos do projeto está disponível na plataforma. [1].

3.2 Projeto e Implementação

3.2.1 Fonte e extração de dados

O Ministério da Saúde disponibiliza todos os dados relativos às mortes em todo o Brasil de 1996 a 2021 por meio da plataforma DataSUS (SAUDE, b), pelo sistema SIM - Sistemas de Informação de Mortalidade (SAUDE, a). A limitação até 2021 convém ao desenvolvimento do projeto, uma vez que a pandemia alterou de maneira excepcional as causas de morte e os seus efeitos ainda são refletidos até o ano de 2023.

A Fiocruz, por sua vez, por meio da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) (FIOCRUZ, b), extrai e enriquece as bases disponibilizadas pelo governo. Isso acontece por meio de sua metodologia ETL (*Extract, Transform and Load*), na qual ela respectivamente acessa e extrai os dados disponíveis e os une no formato adequado, trata os valores e finalmente carrega o resultado dessas operações no sistema do instituto e os disponibiliza à população.

Em particular, na fase de tratamento de dados, muitos valores inválidos são removidos ou tratados, colunas são decodificadas, facilitando a análise. Informações geográficas referentes à localização em coordenadas, a municípios e a unidades federativas são adicionadas, bem como informações relativas ao CID10 (causas de mortes) são adicionadas.

Com esse pré-tratamento, são obtidas bases robustas que podem ser diretamente utilizadas para o estudo. Vale ressaltar que o projeto é atualizado frequentemente pela Fiocruz e a última revisão dele foi realizada no dia 28 de Junho de 2023.

[1] https://github.com/falheisen/mortes

3.2.2 Tratamento da base de dados

O conjunto de dados disponibilizado pela Fiocruz possui 23.5 GB distribuidos em 702 arquivos CSVs (um para cada par ano e estado entre 1996 e 2021). O tamanho dos arquivos é proporcional ao número de mortes de cada estado e, portanto, ao número de instâncias de cada arquivo (cada linha corresponde a um óbito).

O estado de São Paulo (estado mais populoso e sendo assim aquele com o maior número de óbitos anuais), ocupa sozinho 5,52 GB de espaço (6.993.473 instâncias) e leva mais de 1 minuto para ser carregado para manipulação em Python na máquina utilizada. No estado, a base de dados tratada pela Fiocruz apresenta 176 colunas, das quais duas formam um único atributo contador, resultando em 175 atributos. Considerando esse número extremamente elevado de amostras e atributos, dos quais vários eram categóricos, ou seja, seriam decodificados resultando mais colunas, foi necessário realizar uma seleção preliminar de atributos e reduzir o escopo do projeto.

A princípio, todo o Brasil seria avaliado, mas em face dessa limitação, optou-se analisar apenas o estado de São Paulo. Entretanto, foi testado se o modelo construído com essa base conseguiria prever corretamente as instâncias de outros estados, de modo a validar a universalidade da metodologia aplicada e que poderia, assim, ser replicada em outros estados.

3.2.2.1 Seleção preliminar de atributos

Percebeu-se, adicionalmente, que não seria possível aplicar diretamente os métodos de aprendizado de máquina em todos os 175 atributos. Sendo assim, escolheu-se fazer uma seleção preliminar dos atributos. Para isso, primeiramente, selecionaram-se apenas atributos que tinham mais de 60% das amostras com valores não nulos, resultando em 106 atributos. Para reduzir ainda mais a lista de atributos, três casos principais foram explorados:

- Duplicatas de colunas categóricas
- Colunas cuja codificação seria muito complexa
- Colunas que não agregam informação ou com informação irrelevante

O primeiro deles decorre do tratamento da Fiocruz, muitas das colunas estavam "duplicadas", já que contendo a mesma informação, porém, codificadas de maneira diferente. A Tabela 1 exibe o exemplo dos atributos do tipo de óbito.

Em seu tratamento, a Fiocruz "duplicou"todas as colunas com categorias substituindo os códigos pelos textos de cada código. Para a base final as colunas da Fiocruz foram mantidas porque seria mais fácil codificá-las posteriormente.

atributos	Origem	Descrição
TIPOBITO	SIM	Tipo do óbito: 1: óbito fetal 2: óbito não fetal
def_tipo_obito	Fiocruz	Tipo de óbito (Nominal, com as seguintes classificações: Fetal e Não Fetal)

Tabela 1 – Atributos do tipo de óbito

Já no segundo caso, existem colunas que a codificação seria muito complexa, como, por exemplo, o código (ou nome) do município de residência do falecido ou de ocorrência de óbito. Mais especificamente, devido ao fato que o estado de São Paulo contempla 646 municípios, uma codificação única para cada um deles criaria 646 atributos, o que seria impraticável em termos de processamento a ser realizado no escopo do projeto.

Por fim, o terceiro principal caso são de colunas em que não existe informação a ser agregada ao modelo. Exemplos são de atributos em que todos os valores são os mesmos, como é o caso de "TIPOBITO"que tipificam se o óbito é fetal ou não fetal, mas todas as amostras da base são mortes não fetais. Portanto o atributo torna-se irrelevante. Outro caso frequente foram colunas categóricas em que mais de 90% das amostras são "ignorado". Ou seja, não existe informação a ser agregada ao modelo. Isso é o caso das colunas "def_escol_mae", "def_gravidez", ambos com 98% dos valores ignorados.

É válido ressaltar que todos os atributos foram estudados com o auxílio do catálogo da base de dados (FIOCRUZ, a), que reúne 159 atributos. Com isso, foi possível compreender ainda mais os atributos, que foram, então, agregados em algumas características que evidenciam as diferentes análises/causas de mortes. De maneira geral elas podem ser divididas em dois grupos: características do falecido e do óbito em si.

Após a realização de todo o processo descrito anteriormente, foi obtida uma base de dados com 28 atributos, mas muitos dos quais deveriam ser ainda codificados para o modelo.

Ademais, a base conservava as 6.993.473 amostras iniciais, já que nenhuma instância havia sido removida a princípio da base. De modo a permitir a codificação das atributos, todos as amostras que possuíam qualquer valor de atributo inválido foram removidos. Isso fez com que o número de amostras fosse reduzido, nessa etapa, para 5.871.837, o que é um número suficientemente elevado para as etapas subsequentes.

3.2.2.2 Codificação de atributos

Foram feitas codificações de três tipos dos atributos:

- Codificação de datas
- Codificação de OCUP
- Codificação de dados categóricos (exceção OCUP)

Codificação de datas

Este processo envolve a conversão das datas de entrada para um formato padronizado. Neste caso, foi adotado o padrão UNIX, que representa as datas como o número de segundos passados desde 1º de janeiro de 1970, 00:00:00 UTC. Utilizar o padrão UNIX facilita a comparação e o cálculo de intervalos temporais de forma eficiente e consistente.

São atributos que receberam essa codificação: "data_obito"e "data_nasc".

Codificação de OCUP

Esse atributo indica a ocupação do falecido, conforme a Classificação Brasileira de Ocupações (CBO-2002). Essa é uma classificação bem específica e detalhada, o que leva 2658 valores únicos no estado de São Paulo.

Considerando que qualitativamente esse dado poderia ser relevante para o modelo, decidiu-se tratá-lo separadamente para ser adicionado. Para este propósito, foi usado como referência a terceira publicação da classificação (EMPREGO, 2010) que agrega as carreiras por grandes grupos / títulos (10 distintos) e por nível de competência (4 diferentes).

Esses dois atributos foram adicionados ao modelo, substituindo o atributo inicial.

Codificação de dados categóricos

Os dados categóricos restantes, incluindo os dois de ocupação, precisavam ser codificados para ser utilizados no modelo. A escolha foi pelo *One hot encoding*. A razão para isso foi a universalidade de aplicação do método, que não requer um tratamento detalhado prévio de cada atributo. Essa técnica consiste em transformar um atributo categórico com X diferentes valores únicos, em X atributos binários.

3.3 Experimentos

Depois de realizadas as codificações dos atributos, foi obtida uma base com 108 atributos numéricas e uma coluna resposta categórica para 5.871.837 amostras.

A Tabela 2 exibe a relação das causas para as amostras observadas.

Posição	Valor	# Amostras	Proporção
1	IX. Doenças do aparelho circulatório	1.810.604	30,84%
2	II. Neoplasias (tumores)	1.028.392	17,51%
3	X. Doenças do aparelho respiratório	709.942	12,09%
4	XX. Causas externas de morbidade e mortalidade	561.550	9,56%
5	I. Algumas doenças infecciosas e parasitárias	373.800	6,37%
6	XI. Doenças do aparelho digestivo	330.830	5,63%
7	XVIII.Sint sinais e achad anorm ex clín e laborat	321.545	5,48%
8	IV. Doenças endócrinas nutricionais e metabólicas	285.944	4,87%
9	XIV. Doenças do aparelho geniturinário	160.761	2,74%
10	VI. Doenças do sistema nervoso	142.758	2,43%
11	V. Transtornos mentais e comportamentais	50.011	0,85%
12	III. Doenças sangue órgãos hemat e transt imunitár	23.713	0,40%
13	XIII.Doenças sist osteomuscular e tec conjuntivo	23.331	0,40%
14	XII. Doenças da pele e do tecido subcutâneo	20.227	0,34%
15	XVI. Algumas afec originadas no período perinatal	11.019	0,19%
16	XVII.Malf cong deformid e anomalias cromossômicas	10.081	0,17%
17	XV. Gravidez parto e puerpério	6.691	0,11%
18	VIII.Doenças do ouvido e da apófise mastóide	556	0,01%
19	VII. Doenças do olho e anexos	82	0,00%

Tabela 2 – Causas de mortes na base de dados inicial

Existem duas observações relevantes sobre o conjunto de causas: (i) elas são bem numerosas (19) e não são distribuídas proporcionalmente (base desequilibrada).

Como é natural, existem causas que se repetem mais como doenças do aparelho circulatório (doenças cardiovasculares como hipertensão, AVC, etc.), tumores (cânceres) e doenças do aparelho respiratório (asma, pneumonia, etc.). Por outro lado, existem causas que se repetem muito pouco, como doenças do olho e ouvido (menos de 0,1%). Esse desequilíbrio nos dados pode ser um desafio para o modelo, sobretudo nos dados pouco representativos, já que será difícil extrair quais atributos o diferenciam dos demais. Além disso, a redução de classes facilitaria a análise *a posteriori* dos resultados. Em virtude desses pontos, optou-se por agrupar-se as 9 categorias menos representativas, que possuíam cada menos de 1% de representatividade, em uma nova denominada "Outros". A nova relação entre as causas das mortes na base de dados inicial está exibida na Tabela 3.

De modo a equilibrar um pouco a representatividade de cada grupo e acelerar a etapa de testes preliminares, decidiu-se reduzir a base de dados para deixar no máximo 1000 amostras de cada categoria. Sendo assim, obteve-se a relação de causas e amostras exibidas na Tabela 4.

A fim de conduzir os experimentos iniciais na base de teste, escolheu-se utilizar o método XGBoost, detalhado na seção (LI,), dado seu uso extensivo atual em comparações de técnicas de aprendizado de máquina (HOLLMANN SAMUEL MüLLER, 2023).

Posição	Valor	# Amostras	Proporção
1	IX. Doenças do aparelho circulatório	1.810.604	30,84%
2	II. Neoplasias (tumores)	1.028.392	17,51%
3	X. Doenças do aparelho respiratório	709.942	12,09%
4	XX. Causas externas de morbidade e mortalidade	561.550	9,56%
5	I. Algumas doenças infecciosas e parasitárias	373.800	6,37%
6	XI. Doenças do aparelho digestivo	330.830	5,63%
7	XVIII.Sint sinais e achad anorm ex clín e laborat	321.545	5,48%
8	IV. Doenças endócrinas nutricionais e metabólicas	285.944	4,87%
9	XIV. Doenças do aparelho geniturinário	160.761	2,74%
10	Outros	145.711	2,48%
11	VI. Doenças do sistema nervoso	142.758	2,43%

Tabela 3 – Causas de mortes com número de categorias reduzido

Posição	Valor	# Amostras	Proporção
1	IX. Doenças do aparelho circulatório	1.000	9,09%
2	II. Neoplasias (tumores)	1.000	9,09%
3	X. Doenças do aparelho respiratório	1.000	9,09%
4	XX. Causas externas de morbidade e mortalidade	1.000	9,09%
5	I. Algumas doenças infecciosas e parasitárias	1.000	9,09%
6	XI. Doenças do aparelho digestivo	1.000	9,09%
7	XVIII.Sint sinais e achad anorm ex clín e laborat	1.000	9,09%
8	IV. Doenças endócrinas nutricionais e metabólicas	1.000	9,09%
9	XIV. Doenças do aparelho geniturinário	1.000	9,09%
10	Outros	1.000	9,09%
11	VI. Doenças do sistema nervoso	1.000	9,09%

Tabela 4 – Causas de mortes com número de categorias reduzido na base de dados reduzida

3.3.1 Divisão entre dados de treino e teste

Nesse projeto escolheu-se realizar a divisão de dados de treino e teste para o XGBoost usando dois métodos diferentes: (i) *hold-out* e (ii) validação cruzada *k-fold* (*k-fold cross-validation*).

O primeiro método consiste em dividir o conjunto de dados em "treino"e teste", em que o primeiro é utilizado para treinar o modelo (minimizar alguma função) e o segundo é usado para avaliar a performance do modelo em novas amostras. Um valor comum e adotado nesse projeto é 70% das amostras para treino e 20% para testes. Decidiu-se utilizar estratificação para garantir que o número de casos de teste e treino seja o mesmo para cada categoria.

O segundo método consiste em dividir o conjunto de dados em k grupos. Um dos conjuntos é usado como teste e os *k-1* demais são usados para treinamento. O modelo é treinado nos conjuntos de treinamento e a performance é avaliada no conjunto de teste. O processo é repetido k vezes, até que todos os grupos sejam usados como conjunto de teste.

A vantagem desse método é permitir a avaliação robusta do desempenho do modelo em diferentes conjuntos de dados, mitigando possíveis viéses associados a uma única divisão de treino e teste. Adotou-se arbitrariamente 5 folds para o estudo.

A ideia de contar com as duas alternativas era avaliar o impacto das duas, bem como permitir a calibragem de hiperparâmetros (no caso da validação cruzada).

3.3.2 Calibragem de hiperparâmetros

Utilizando a divisão do conjunto de dados *5-fold cross validation*, escolheu-se a otimização bayesiana (*CV Bayes Cross Validation*), a fim de calibrar os hiperparâmetros do modelo.

Nessa etapa, foi necessário utilizar um número de iterações igual a 200, já que valores inferiores, como 50, não permitiam a calibragem adequada. Apesar de ser um método eficiente, implicou em custo computacional elevado quase três horas de custo computacional com um processo rodando em paralelo para cada núcleo de processamento da máquina.

As faixas dos atributos a serem testadas estão exibidas na Tabela e os valores obtidos para os hiperparâmetros para o XGBoost está exibido na Tabela 5.

atributo	Tipo	Valor testado inicial	Valor testado final	Valor obtido
Taxa de aprendizado	Float	1e-7	1	0.64
Profundidade máxima	Int	1	10	6
Subamostra	Float	0,2	1	1
Subamostra por árvore	Float	0,2	1	0.2
Peso mínimo por filho	Float	1e-16	1e5	1e-16
Alpha	Float	1e-16	1e2	15.86
Lambda	Float	1e-16	1e2	100
Gamma	Float	1e-16	1e2	1e-16
# Estimadores	Int	100	4000	100

Tabela 5 – Valores dos hiperparâmetros para o método XGBoost, utilizando *5-fold cross-validation* e otimização dos hiperparâmetros por Bayes

3.3.3 Resultados preliminares

Em cada um dos *folds* foi utilizado o modelo, ou seja, a hiperparametrização obtida por meio da calibragem. A acurácia média nesses testes foi de 31,02% ($\pm$0.0131)% e o melhor resultado foi de 32,36...% Apesar de parecer muito baixo, esse valor é muito superior a uma predição aleatória de 9,09...% (1/11, e que 11 corresponde ao número de categorias). Isso acontece porque no problema em questão há um número muito elevado de classes.

A matriz de confusão do caso de maior acurácia está exibida na Tabela 6. Para exibição nesse trabalho, nas linhas estão as classes reais dos dados de teste, enquanto nas colunas estão as classes preditas, identificadas com os rótulos numéricos que correspondem às classes.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**	**Acurácia**
I. Algumas doenças infecciosas e parasitárias	**106**	12	13	5	5	4	5	15	13	6	16	200	**53%**
II. Neoplasias (tumores)	23	**70**	15	14	4	10	6	17	20	11	7	197	**36%**
IV. Doenças endócrinas nutricionais e metabólicas	21	32	**22**	12	5	14	18	16	20	25	3	188	**12%**
IX. Doenças do aparelho circulatório	9	23	15	**39**	3	13	22	16	20	29	8	197	**20%**
Outros	35	11	13	10	**20**	16	19	21	20	20	12	197	**10%**
VI. Doenças do sistema nervoso	15	18	15	8	11	**48**	8	12	34	16	17	202	**24%**
X. Doenças do aparelho respiratório	20	22	19	18	6	11	**37**	8	37	14	13	205	**18%**
XI. Doenças do aparelho digestivo	14	37	21	19	5	10	11	**52**	17	13	19	218	**24%**
XIV. Doenças do aparelho geniturinário	16	22	19	13	2	25	27	10	**48**	11	1	194	**25%**
XVIII.Sint sinais e achad anorm ex clín e laborat	7	12	3	4	2	12	2	7	8	**116**	18	191	**61%**
XX. Causas externas de morbidade e mortalidade	5	5	2	8	3	6	2	11	5	10	**154**	211	**73%**
Total	271	264	157	150	66	169	157	185	242	271	268		**32%**

Tabela 6 – Matriz de confusão obtida para o modelo XGBoost, utilizando *5-fold cross-validation* e otimização dos hiperparâmetros por Bayes

A análise da matriz revela que a acurácia das categorias varia muito. Enquanto há categorias como "XX. Causas externas de morbidade e mortalidade", que apresenta acurácia acima de 70%, há categorias como "IV. Doenças endócrinas nutricionais e metabólicas"(e.g. diabetes, hipotireodismo...), que apresenta uma acurácia de apenas 12%. O destaque negativo fica evidente para a classe "Outros", que apresenta apenas um ponto percentual de acurácia acima do resultado aleatório. Isso sugere que essa redução do número de classes pode ter prejudicado o reconhecimento das causas de morte agrupadas.

Em relação à semelhança de algumas classes e dificuldade de reconhecimento no modelo, as amostras da categoria "IV. Doenças endócrinas nutricionais e metabólicas"foi predita mais vezes como sendo das categorias "II. Neoplasias (tumores)"e "XVIII.Sint sinais e achad anorm ex clín e laborat"do que da classe correta. Isso indica que o modelo não é muito confiável para identificar amostras dessa classe. Transferindo para o cenário prático, a matriz de confusão indicaria que um falecido por conta de diabetes ou hipotireoidismo, por exemplo, teria uma probabilidade maior de ser classificado na categoria de tumores (câncer) do que na categoria apropriada. Não é possível por meio dessa análise identificar se existe uma doença específica que causaria esse problema na assertividade do modelo e esse ponto está além do escopo de pesquisa deste projeto.

Um fenômeno semelhante acontece também com as classes "Outros"e "X. Doenças do aparelho respiratório". A Tabela 7 reexibe a matriz de confusão, destacando classes que foram preditas em um número igual ou superior da classe correta, incorrendo no problema acima.

Após de obtidos os resultados acima do modelo com XGBoost realizado com dados de treino e teste divididos por meio do 5-cross validation com calibragem de hiperparâmetros pelo método de Bayes, foi testado o resultado obtido para o mesmo modelo, com dados de

	0	1	2	3	4	5	6	7	8	9	10	Total	Acurácia
I. Algumas doenças infecciosas e parasitárias	**106**	12	13	5	5	4	5	15	13	6	16	200	**53%**
II. Neoplasias (tumores)	23	**70**	15	14	4	10	6	17	20	11	7	197	**36%**
IV. Doenças endócrinas nutricionais e metabólicas	21	32	**22**	12	5	14	18	16	20	25	3	188	**12%**
IX. Doenças do aparelho circulatório	9	23	15	**39**	3	13	22	16	20	29	8	197	**20%**
Outros	35	11	13	10	**20**	16	19	21	20	20	12	197	**10%**
VI. Doenças do sistema nervoso	15	18	15	8	11	**48**	8	12	34	16	17	202	**24%**
X. Doenças do aparelho respiratório	20	22	19	18	6	11	**37**	8	37	14	13	205	**18%**
XI. Doenças do aparelho digestivo	14	37	21	19	5	10	11	**52**	17	13	19	218	**24%**
XIV. Doenças do aparelho geniturinário	16	22	19	13	2	25	27	10	**48**	11	1	194	**25%**
XVIII.Sint sinais e achad anorm ex clín e laborat	7	12	3	4	2	12	2	7	8	**116**	18	191	**61%**
XX. Causas externas de morbidade e mortalidade	5	5	2	8	3	6	2	11	5	10	**154**	211	**73%**
Total	271	264	157	150	66	169	157	185	242	271	268		**32%**

Tabela 7 – Matriz de confusão obtida com destaque para as classes incorretas que foram preditas em um número igual ou superior da classe correta

treino e teste pelo método hold-out.

A acurácia obtida foi de 28,45%, ou seja 3,57 pontos percentuais abaixo do valor médio obtido previamente (ou aproximadamente 10% inferior). Enquanto a acurária top-3 diferiu de apenas 2,34 pontos percentuais. Por outro lado, o custo computacional de calibrar os hiperparâmetros é consideravelmente menor. A Tabela 8 exibe a relação entre os valores e a Figura 1 exibe o grau de acurácia da predição com base no valor de k adotado.

Modelo	Acurácia Regular	Acurácia Top-3	Custo computacional (segundos)
5-Cross Validation com calibragem de hiperparâmetros (valores médios)	31,02%	56,27%	8640
Hold-out e estratificação das classes	28,45%	53,27%	<2

Tabela 8 – Comparação de desempenho dos métodos para o método XGBoost

A Tabela 9 exibe a matriz de confusão obtida e a comparação da acurácia para a precisão de cada categoria ao lado. É importante ressaltar que a comparação é em relação ao melhor caso dos 5 modelos gerados previamente (em relação à precisão). É possível notar que há piora na acurácia de todas as categorias, com exceção da categoria "Outros", em que há melhoria significativa da predição. No entanto, a diferença não parece acentuada em quase nenhum categoria, exceto para "II. Neoplasias (tumores)"e mesmo esse sendo o melhor caso de precisão da abordagem mais complexa, a diferença total é ainda abaixo de 4%.

Como no outro caso, "IV. Doenças endócrinas nutricionais e metabólicas"e "X. Doenças do aparelho respiratório"têm o problema de que outras classes foram indicadas o mesmo número de vezes (ou maior) que elas. No entanto, "Outros"não tem mais esse problema.

Considerando que o objetivo do projeto consiste em explorar diversos modelos diferentes, alguns dos quais a calibragem é computacionalmente mais custosa que o XGBoost (como o Catboost), julgou-se que o custo computacional da realização das

	0	1	2	3	4	5	6	7	8	9	10	Total	Acurácia	Variação Acurácia
I. Algumas doenças infecciosas e parasitárias	**134**	28	10	10	21	19	11	26	20	10	11	300	**45%**	-8%
II. Neoplasias (tumores)	21	**65**	30	37	14	19	27	36	30	17	4	300	**22%**	-14%
IV. Doenças endócrinas nutricionais e metabólicas	32	38	**36**	26	15	21	29	36	32	30	5	300	**12%**	0%
IX. Doenças do aparelho circulatório	20	35	29	**51**	17	22	28	26	27	32	13	300	**17%**	-3%
Outros	24	27	16	17	**67**	22	32	28	30	24	13	300	**22%**	12%
VI. Doenças do sistema nervoso	31	20	25	19	32	**55**	28	19	45	20	6	300	**18%**	-5%
X. Doenças do aparelho respiratório	21	30	32	**40**	17	27	**40**	29	38	14	12	300	**13%**	-5%
XI. Doenças do aparelho digestivo	24	44	25	35	15	14	17	**53**	36	17	20	300	**18%**	-6%
XIV. Doenças do aparelho geniturinário	42	35	20	19	17	37	36	21	**52**	14	7	300	**17%**	-7%
XVIII.Sint sinais e achad anorm ex clín e laborat	11	16	16	11	12	14	11	8	14	**167**	20	300	**56%**	-5%
XX. Causas externas de morbidade e mortalidade	5	10	0	10	6	6	11	16	2	15	**219**	300	**73%**	0%
Total	365	348	239	275	233	256	270	298	326	360	330	3300	**28%**	3,91%

Tabela 9 – Matriz de confusão obtida para o método XGBoost, utilizando hold-out e estratificação das classes (sem calibragem dos hiperparâmetros)

etapas de cross-validation e calibragem dos hiperparâmetros não justificaria o aumento da acurácia. Por conta disso, decidiu-se abandonar essa estratégia para o estudo comparativo dos diferentes modelos. Sendo assim, doravante adotou-se o método hold-out e não se calibrou o modelo antes de utilizá-lo.

Adotando-se esse modelo, é possível concluir que, por meio do conjunto de dados estudado, que reflete particularidade do paciente e óbito, é possível predizer com certa confiabilidade a morte de um falecido. Tais resultados, portanto, respondem nossa primeira questão de pesquisa.

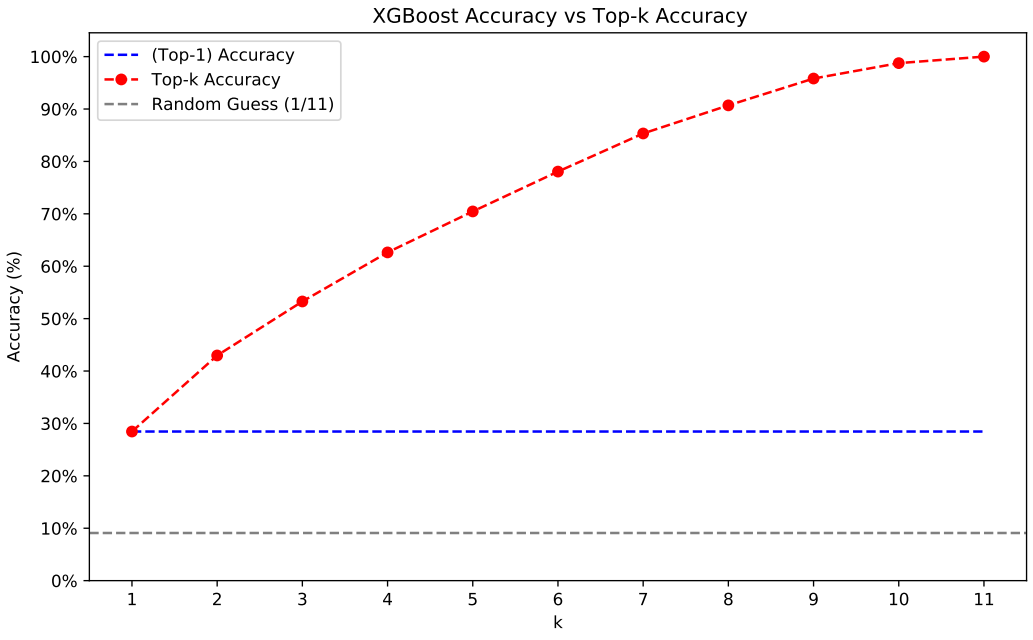

Figura 1 – Acurácia da previsão para valores distintos de k

3.3.4 Comparação de diferentes modelos

A fim de encontrar um modelo que maximizasse a precisão da predição de classes (ainda na primeira questão de pesquisa) e respondesse a segunda questão de pesquisa de generalidade do modelo, buscou-se aplicar modelos diferentes de aprendizado supervisionado no conjunto de dados.

Em particular, foram aplicados métodos baseados em boosting e em redes neurais profundas.

Técnicas baseadas em boosting

Foram aplicadas duas técnicas baseadas em boosting adicionais ao XGBoost, o Catboost e o LightGBM. A Tabela 10 exibe a relação das acurácias obtidas.

Modelo	Acurácia Regular	Acurácia Top-3
XGBoost	28.45%	53.27%
LightGBM	29.82%	54.24%
CatBoost	30.24%	54.79%

Tabela 10 – Acurácia dos métodos preditivos baseados em Boosting

Nota-se que a acurácia obtida é bem similar para os três métodos, tanto para k=1 (acurácia normal), quanto para k=3, na visão top-k. Apesar disso, existe uma relação em que o XGBoost tem a menor acurácia para os dois valores de k, seguido pelo LightGBM e por fim pelo CatBoost, para valores de k pequenos ($k < 6$). Em termos de custo computacional, XGBoost e LightGBM são muito superiores ao CatBoost que demora a ser executado na configuração experimental do estudo.

A Figura 2 exibe a acurácia para diferentes valores de de k. A imagem reforça a ideia de que os modelos tem valores muito próximos, agora inclusive para todos os valores de k.

Técnicas baseadas em redes neurais profundas

Foram aplicadas duas técnicas baseadas em redes neurais profundas: o MLP Residual (TOUVRON et al., 2023) e o Transformer (VASWANI et al., 2017). A Tabela 11 exibe a relação das acurácias obtidas, com o XGBoost como referência dos métodos baseados em boosting.

Modelo	Acurácia Regular	Acurácia Top-3
XGBoost	28.45%	53.27%
MLP Residual	21.06%	44.97%
Transformer	22.58%	46.55%

Tabela 11 – Acurácia dos métodos preditivos baseados em Boosting

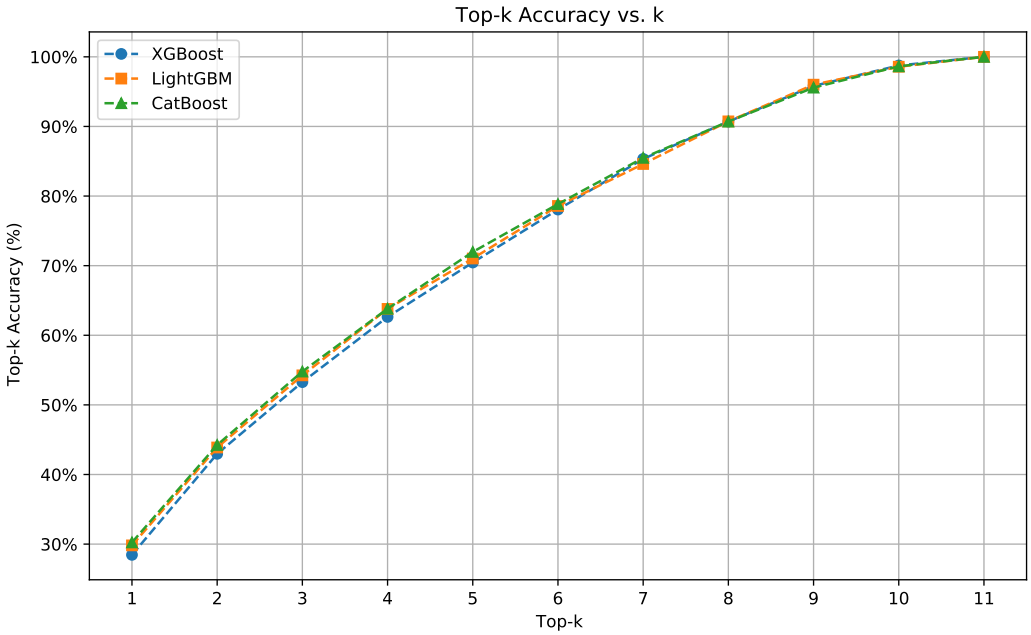

Figura 2 – Acurácia da previsão para valores distintos de k

É possível observar uma grande diferença de acurácia dos métodos baseados em redes neurais profundas com o XGBoost tanto para k=1, quanto para k=3. A Figura 3 exibe a relação da acurácia para diferentes valores de k.

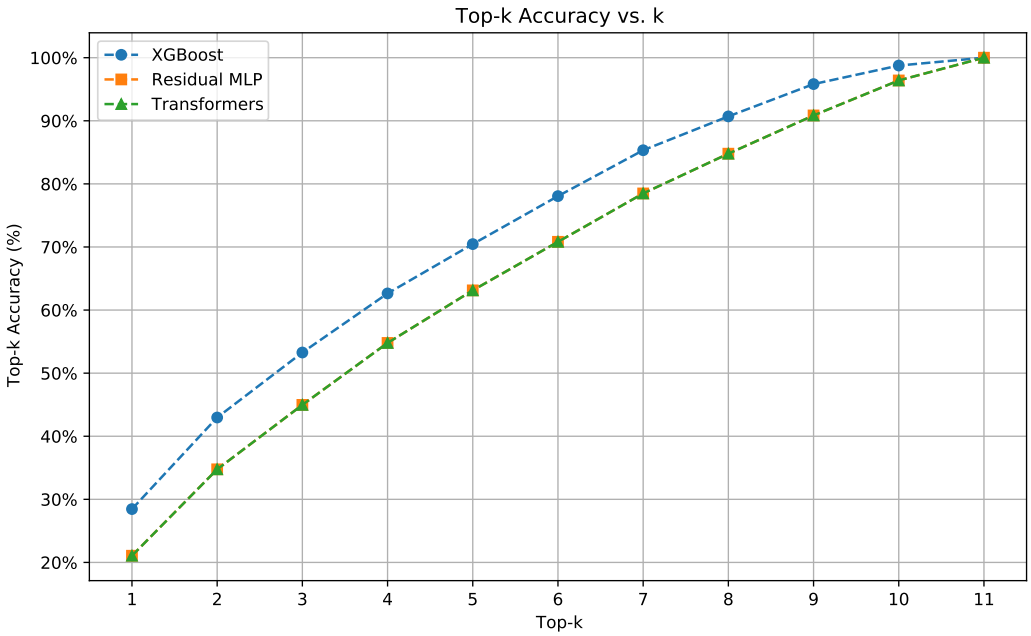

Figura 3 – Acurácia da previsão para valores distintos de k

É possível concluir que os métodos baseados em *boosting* são melhores, quando aplicados no conjunto de dados, com a metodologia adotada, isto é de estratificar as classes e não realizar uma calibração específica nos modelos antes de usá-los. Independentemente da razão, os métodos de boosting parecem mais promissores no contexto do estudo.

Ainda, pela proximidade da precisão dessas técnicas, elas parecem igualmente promissoras para explorar a generalização do modelo. Por conta do custo computacional mais elevado, o CatBoost foi despriorizado. Para decidir entre o LightGBM, considerou-se a maior abrangência e conhecimento do XGBoost, que inclusive foi o mesmo motivo para começar o estudo com ele.

3.3.5 Seleção de atributos

Uma vez selecionada a técnica a ser utilizada doravante, a próxima etapa a ser realizada era fazer uma seleção de atributos. Isto é, entender o impacto de cada uma na predição do modelo. Para isso foram utilizados dois algoritmos de estado da arte, os *Infinite Attributes Selecion (IFS)* supervisionados e não supervisionados (Inf-FS_S e Inf-FS_U respectivamente) (ROFFO SIMONE MELZI, 2020).

A Tabela 12 exibe as atributos que são as 10% mais relevantes identificadas pelos métodos, com a posição de cada uma em cada abordagem.

atributo	Descrição	Top 10% FS_S	Top 10% FS_U	Posição FS_S	Posição FS_U
data_nasc	Data de nascimento	1	1	1	1
ano_nasc	Ano do nascimento	1	1	2	5
data_obito	Data de ocorrência do óbito	1	1	3	2
def_cirurgia_sim	Indica que houve cirurgia (categoria)	1	0	4	39
ano_obito	Ano do óbito	1	0	5	16
idade_obito_anos	Idade do óbito (em anos) informada na declaração de óbito	1	1	6	6
idade_obito_calculado	Idade do óbito calculado utilizando a data de óbito e a data denascimento	1	0	7	14
idade_obito	Idade do óbito reportada	1	0	8	11
def_loc_ocor_domicílio	Local de ocorrência do óbito no domicílio (categoria)	1	0	9	21
def_exame_ignorado	É ignorada a realização de exame (categoria)	1	0	10	36
dia_semana_obito_sab	Óbito ocorreu no sábado (categoria)	0	1	98	3
dia_semana_nasc_qua	Nascimento ocorreu na quarta-feira (categoria)	0	1	57	4
res_latitude	Latitude da sede do Município de residência da pessoa que foi à óbito	0	1	54	7
res_longitude	Longitude da sede do Município de residência da pessoa que foi à óbito	0	1	79	8
dia_semana_obito_qua	Óbito ocorreu na quarta (categoria)	0	1	95	9
dia_semana_obito_dom	Óbito ocorreu no domingo (categoria)	0	1	42	10

Tabela 12 – Top 10% das atributos mais relevantes do modelo

A análise da tabela revela que existem apenas 4 atributos comuns no top-10 dos dois métodos. Além disso, as três ocupam posições "baixas"(no máximo sexto) e duas são a primeira e a segunda nos dois. Todas elas são relacionadas à idade da pessoa que faleceu. Isso indica que esse é a dimensão mais importante para a causa da morte.

Em particular, a atributo "data_nasc"se mostrou a mais relevante para os dois métodos. A Figura 4 exibe um gráfico de violino dessa atributo para cada categoria, em que o eixo y apresenta o valor com a codificação UNIX do nascimento. É possível visualizar como de fato algumas categorias possuem valores que se distanciam de outras, como a categoria ’0’ ("I. Algumas doenças infecciosas e parasitárias") e 10 ("XX. Causas externas

de morbidade e mortalidade"), que obtiveram precisões bem elevadas (respectivamente 45% e 73%).

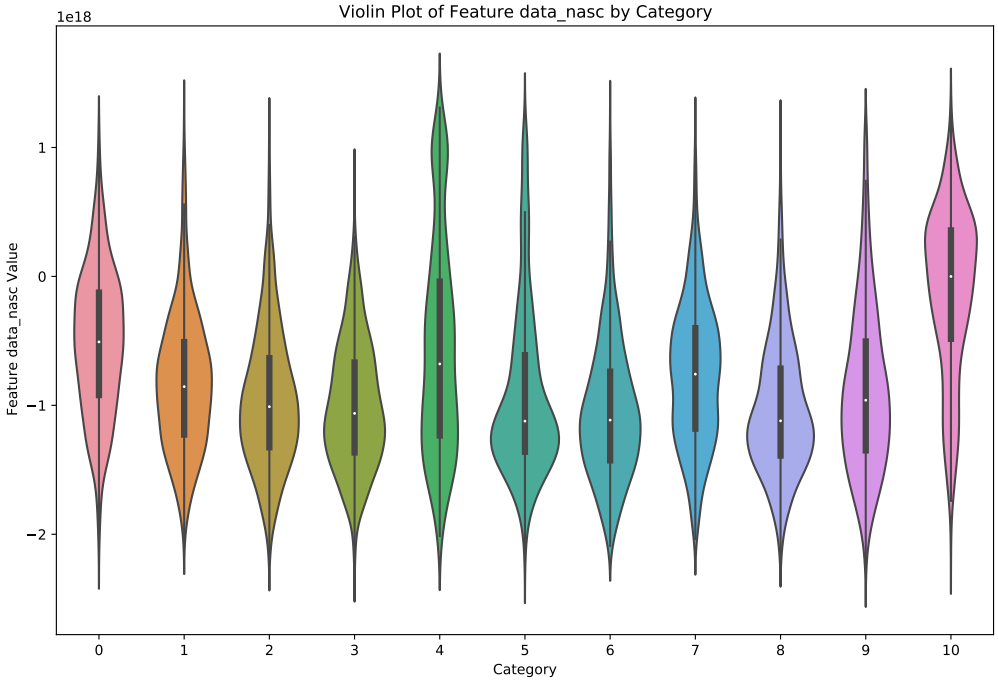

Figura 4 – Valor da atributo "data_nasc"para cada categoria

O valor superior da atributo implica que na verdade as pessoas nessas categorias tem uma morte mais prematura que em relação as outras categorias. Intuitivamente, faz sentido que as pessoas que morram mais cedo que a média seja por causas externas de morbidade e mortalidade que não as doenças mapeadas pelas outras categorias.

Ainda com base na Tabela, o método supervisionado identificou outras dimensões (além da idade do falecido) que se destacam na identificação da causa da morte na análise, informações relacionadas à crise que levou o indivíduo a óbito como o fato que aconteceu cirurgia, que a crise aconteceu em casa e se exames foram realizados.

Por outro lado, o método não supervisionado identifica as dimensões geográficas, que provavelmente reflete fatores socio-econômicos relacionados aos municípios e uma questão de datas. Nas categorias que indicam que o óbito aconteceu na quarta e domingo pode revelar uma variação da rotina de atendimento nessas datas. Questões como disponibilidade de médicos para urgências, janelas de plantões e outros aspectos podem refletir-se nos dados. Pode, também, indicar que o paciente dá entrada no hospital dias antes e o óbito ocorre nessas datas por questões naturais. Supõe-se, por exemplo, que os pacientes com desconfortos esperem até o sábado para irem ao hospital. Nesse ambiente descobrem um problema que deveria ter sido tratado antes e acabam indo ao óbito no domingo.

Existe apenas uma atributo que surgiu no método não supervisionado pra qual é mais difícil levantar hipóteses que justifiquem sua importância: a que indica que o nascimento do falecido ocorreu na quarta-feira. Para essa não especulam-se razões para importância.

A fim de determinar-se qual a relevância dos 10% das atributos selecionadas pelos métodos e entender qual a importância de estender-se o número de atributos, foi estimada a acurácia (top-1) e top-3 do modelo XGBoost construído.

A Figura 5 apresenta os valores obtidos.

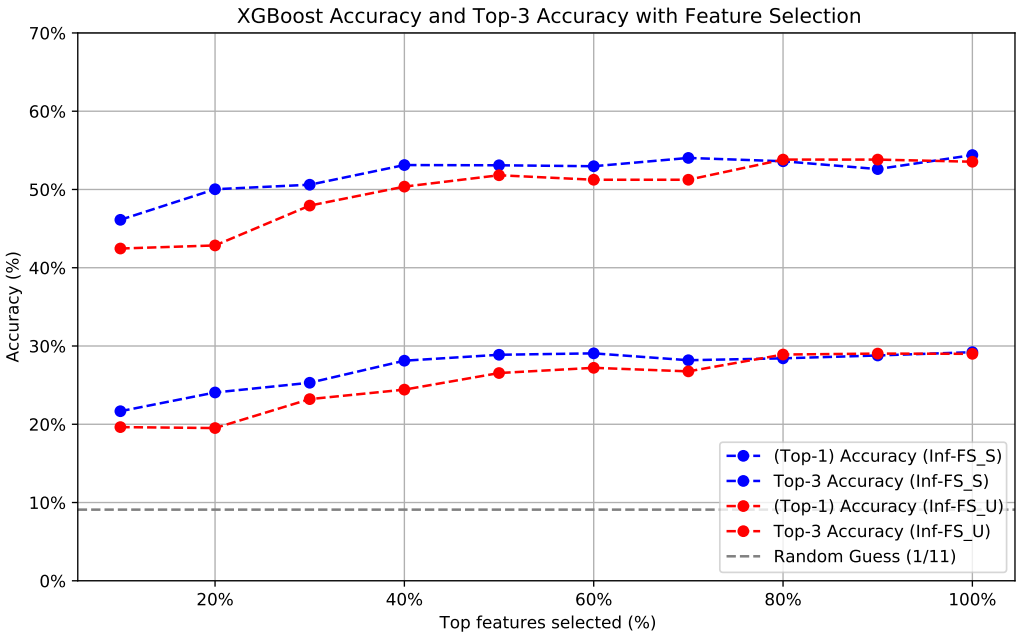

Figura 5 – Impacto da seleção de atributos sobre a acurácia do modelo

Com base nessas imagens, nota-se que a acurácia varia pouco (aproximadamente 10 pontos percentuais) tanto para a top-1, quanto para a top-3 de 10 a 100% das atributos (10 a 108 atributos), em ambos os métodos, com uma performance melhor para a seleção de atributos pelo método supervisionado.

Nesse método, com cerca de 40% dos atributos (43 de 108 atributos) a acurácia praticamente não se altera para a acurácia top-1 e top-3. Depois dessa fração, a acurácia top-3 chega a subir com 70%, mas esse crescimento é reduzido pela queda de acurácia em 90%. Para acurácia regular, o valor máximo é obtido com 60%, cai com 70% e é retomado aos pontos até 100%.

Ainda que o método não supervisionado se mostre levemente inferior nesse caso de estudo, a diferença de acurácia top-1 e top-3 com apenas 50% dos atributos é muito próximo do valor com todos. Nesse método, não há variação entre 10 e 20% dos atributos,

razão pela qual nesta porcentagem há a maior diferença de acurácia tanto top-1 quanto top-3 para os métodos.

Com base em todos esses pontos, é possível concluir que existem atributos que desempenham maior impacto na predição da causa da morte. Esses resultados sugerem que a resposta para a terceira questão de pesquisa deste projeto é positiva.

3.3.6 Aplicação do modelo a novas amostras

Um ponto crucial para avaliação do modelo é determinar como ele generaliza a amostras externas aos casos de treino e teste. Nesse estudo a relevância dessa etapa é ainda maior, considerando que apenas 11000 amostras de mais de 5 milhões de amostras do estado de São Paulo foram utilizadas nas etapas de teste e treino.

Ademais, apenas foram utilizados dados do Estado de São Paulo, quando a ideia é que as predições do modelo sejam aplicáveis para todo o país. Por último, os dados foram treinados e testados em uma base equilibrada, isto é, com um número de amostras semelhante em cada classe, quando na verdade as amostras na vida prática são desequilibradas.

Todos esses fatores combinados indicam que talvez o modelo não performe bem em amostras que externas ao conjunto dos dados de treino e teste.

A fim de verificar essa premissa, o modelo teria que ser alterado. As atributos 'res_MSAUDCOD', 'res_RSAUDCOD', 'res_CSAUDCOD' estavam presentes no modelo original (abordado precedentemente). Elas são códigos referentes respectivamente à macrorregional, a regional e a microrregional a que o falecido residia. Por isso, quando foram codificadas se tornaram atributos que estariam presentes apenas no estado de São Paulo. Além disso, quando fossem decodificadas em amostras de outros estados gerariam atributos incompatíveis. Sem esses 3 atributos originais, após a codificação as 108 atributos foram reduzidas para 80. Ou seja, 28 foram eliminadas do modelo. A Tabela 13 exibe a nova acurácia obtida para o modelo sem as 28 atributos.

Modelo	Acurácia Regular	Acurácia Top-3
XGBoost	29.12%	53.82%
LightGBM	29.85%	54.42%
CatBoost	29.36%	54.42%

Tabela 13 – Acurácia dos métodos preditivos baseados em Boosting

Nota-se que praticamente não há diferença nas acurácias obtidas com as atributos anteriores e com as atuais. Ainda que a acurácia do XGBoost e LightGBM aumente tanto a para o top-1 quanto top-3 e que, por outro lado, o Catboost reduza suas acurácias, a diferença é de menos 1 ponto percentual em todos os casos.

Isso mostra como as 28 atributos (26% do total) poderiam ser excluídas do modelo, como foi possível observar por meio da análise de seleção de atributos.

Pelos mesmos motivos discutidos anteriormente, decidiu-se aplicar o XGBoost para as novas amostras. Para selecionar as amostras às quais o modelo seria aplicado, selecionou-se os estados mais ricos de cada região: São Paulo (Sudeste), Rio Grande do Sul (Sul), Goiás (Centro-Oeste), Pará (Norte), com exceção apenas da região Nordeste para qual selecionou-se o estado de Pernambuco. Isso foi feito porque o Estado da Bahia não possuía uma atributo presente no modelo.

A Tabela 14 apresenta as acurácias obtidas nos ensaios.

Estado	Acurácia Regular	Acurácia Top-3
Predição aleatória	9.09%	-
São Paulo*	26.33%	51.67%
Rio Grande do Sul	16.43%	37.37%
Pernambuco	29.19%	44.27%
Goiás	21.26%	45.93%
Pará	27.85%	50.38%

Tabela 14 – Aplicação do modelo para diferentes estados

A fim de aferir a acurácia no estado de São Paulo, seria conceitualmente apropriado remover os dados de treino e teste do conjunto de dados. No entanto, julgou-se que isso não seria relevante, já que apenas 11.000 dos mais de 5.5 milhões de amostras do estado se encaixam nessa categoria. Esse ponto justifica o asterisco do estado na Tabela.

Observa-se que o Rio Grande do Sul é o estado no qual o modelo apresenta a pior performance preditiva. Isso é um dado interessante, considerando que os dados foram testados no estado de São Paulo e o Rio Grande do Sul é o estado que tem a expectativa de vida mais próxima do estado de São Paulo dentre todos os outros estados para o ano de 2022 (SIDRA. . . ,) (79,54 para SP e 79,26 para RS). Além disso, em outros termos econômicos e sociais o estado também é o que mais se aproxima de São Paulo.

A Figura 6 apresenta a evolução da acurácia com o aumento do k. Ela evidencia ainda mais a falta de performance das amostras do RS em relação aos demais estados.

De maneira igualmente surpreendente, no estado do Pará, o modelo apresenta resultados de acurácia top-1 relevantes, com valores superiores até aos do próprio estado de São Paulo. Considerando a distância dos estados em indicadores econômicos, sociais e de saúde, como a própria expectativa de vida que no Pará é de 73,18 anos: mais de 6 anos de diferença em relação à SP.

Outro destaque positivo vai para o estado de Pernambuco que registra a maior acurácia top-1: de 29,19%, quase 3 pontos percentuais acima do estado de São Paulo. Apesar disso, sua acurácia top-3 é inferior a do Para e a de São Paulo.

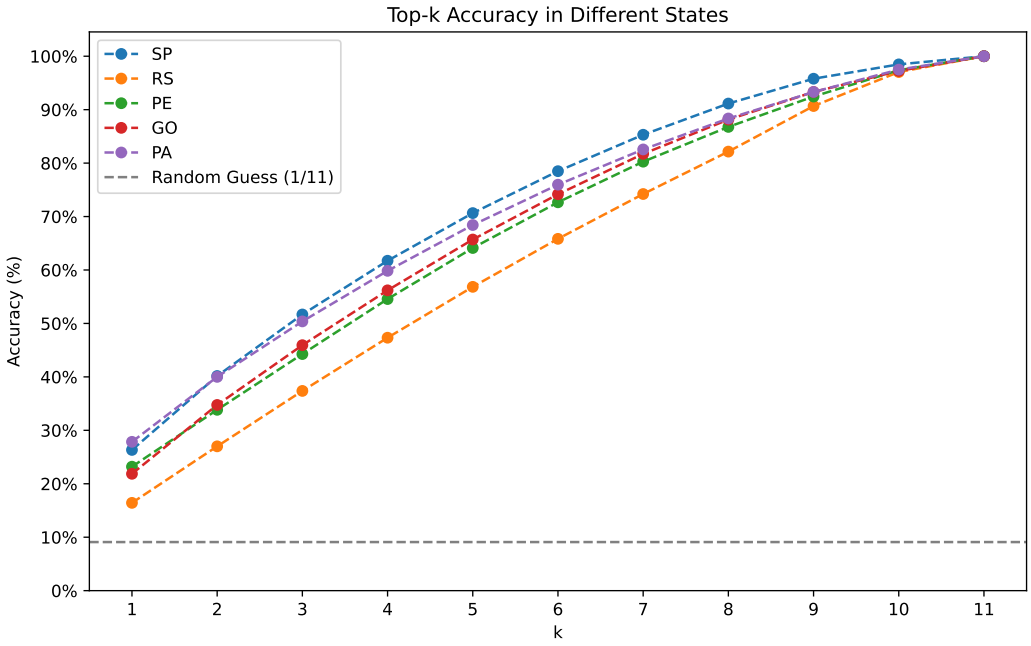

Figura 6 – Acurácia do modelo aplicado em diferentes estados para diferentes valores de k

São Paulo apresenta uma queda de acurácia top-1 quase 3 pontos percentuais abaixo do conjunto de testes, abaixo de dois outros estados e acima de outros dois, nessa análise. Na acurácia top-3 ele possui a maior acurácia, cerca de 2 pontos percentuais abaixo do modelo aplicado pro conjunto de testes. A queda ainda distancia o resultado, não obstante, do palpite aleatório.

A aplicação do modelo nessas amostras permite concluir que os modelos desenvolvidos generalizam os resultados obtidos para amostras externas aos casos de treinamento e teste, inclusive mantém sua capacidade preditiva para outros estados. Portanto, encontra-se uma resposta positiva para a segunda questão de pesquisa.

4 Considerações Finais

Ao longo do desenvolvimento deste projeto foi possível predizer a causa de morte de um falecido a partir de um conjunto de dados públicos no Brasil, que reflita particularidades do paciente e do óbito registrado. Além disso, verificou-se que os modelos generalizavam os resultados obtidos para amostras externas aos casos de treinamento e teste, apenas de São Paulo, para amostras de diferentes estados da federação, com características socioeconômicas que diferem muito de SP. Finalmente foi possível identificar uma relação de features que possui maior impacto na predição da causa da morte.

Sendo assim, todas as questões de pesquisa propostas no estudo foram respondidas de forma positiva. Com isso é possível concluir que o trabalho foi concluído com sucesso, isto é (i) que é possível predizer a causa da morte de um indivíduo, a partir de um conjunto de dados que reflita particularidades do paciente e do óbito registrado, (ii) que os modelos desenvolvidos generalizaram as predições para amostras aos casos de treinamento e teste, para todo país e (iii) que existem atributos que desempenham maior impacto na predição da causa da morte.

4.1 Perspectivas de Continuidade

Ao longo desse estudo, foram tomadas decisões em diferentes momentos que deixaram questões em aberto que podem ser revisadas em estudos posteriores.

Um deles é em relação a estudo dos dados até 2021. Com isso foi deixado de lado os anos mais recentes impactados pela pandemia de 2019. Seria interessante entender como o modelo desenvolvido se comportaria nos anos seguintes e se o padrão de mortalidade se manteve, ou houve uma ruptura de hábitos e na saúde pública de modo que não é possível utilizar um mesmo modelo.

Outro ponto é em relação a quantidade massiva de dados. Em virtude dela, foi escolhido fazer um recorte balanceado de categorias e amostras para treinar e testar o modelo. Seria interessante revisitar o estudo com novas estratégias de redução de custo computacional ou uma disponibilidade maior de custo computacional a fim de utilizar uma quantidade mais significativa de dados e avaliar a importância disso no modelo. Ademais, seria também relevante observar o impacto do recorte ter sido balanceado na capacidade preditiva do modelo. Ainda nesse tópico, essas estratégias poderiam ser importantes não apenas para revisitar o estudo, mas para desenvolver uma estratégia de acompanhamento de um modelo preditivo, considerando que a tendência é que o número de dados apenas aumente com os anos que virão.

Uma terceira questão que pode ser enriquecida é a engenharia de features do projeto. Foi utilizada a base da Fiocruz para desenvolvimento do projeto, sem anexar nenhum indicador adicional. Seria por exemplo possível adicionar a renda per capta do município em que vive o falecido, entre diversos outros indicadores socioeconômicos à disposição. Além disso, por questões relacionadas ao contexto do projeto, a seleção preliminar de features excluiu atributos que poderiam ter relevância para um modelo preditivo. Uma revisitação dessa etapa poderia ser igualmente frutífera. Ademais, a codificação das features, sejam elas adicionais ou não, poderia ser também revista. Foram utilizadas apenas 3 codificações diferentes para as features, quando é possível que teriam técnicas mais adequadas para determinadas.

Uma quarta questão diz respeito a criação da categoria "Outros". Seria interessante avaliar o impacto dessa medida na capacidade preditiva do modelo.

Uma quinta questão seria estudar o impacto da parametrização dos modelos de redes neurais profundas. No estudo foi possível observar que o XGBoost não teve alteração signficativa da performance quando calibrado. No entanto, isso não necessariamente é verdade para as técnicas de redes neurais profundas, mas que foi assumido por limitações de custo computacional no projeto.

Outra perspectiva de continuidade não diz respeito à realização de novos testes, mas a re-interpretação dos testes já existentes. O estudo conduzido com foco e conhecimento maior nos modelos e técnicas de aprendizado de máquina a serem implementados. Uma revisão do projeto conduzida por um profissional da área de saúde poderia trazer novas interpretações e perspectiva que enriqueceriam os pontos observados, bem como identificar novas vias de desenvolvimento do projeto.

Referências

BATISTA, L. E.; ESCUDER, M. M. L.; PEREIRA, J. C. R. A cor da morte: causas de óbito segundo características de raça no estado de são paulo, 1999 a 2001. *Rev Saúde Pública*, 2004. Citado na página 17.

BURKOV, A. *The Hundred-Page Machine Learning Book*. [S.l.]: McGraw-Hill Science/Engineering/Math, 2019. Citado na página 21.

CHANG, M.-H. et al. Trends in disparity by sex and race/ethnicity for the leading causes of death in the united states-1999-2010. *J Public Health Manag Pract.*, 2016. Citado na página 17.

EMPREGO, M. do Trabalho e. *Classificação Brasileira de Ocupações*. 3a. ed. Brasília: Ceira - Coimbra, 2010. Citado na página 32.

FIOCRUZ. *Dicionário de variáveis do SIM*. Rio de Janeiro: [s.n.]. <https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacoes-de-mortalidade-sim/dicionario-de-variaveis/>. Acesso em: 10 set 2023. Citado na página 31.

FIOCRUZ. *Plataforma de Ciência de Dados Aplicada à Saúde (PCDaS)*. Rio de Janeiro: [s.n.]. <https://pcdas.icict.fiocruz.br/>. Acesso em: 10 set 2023. Citado na página 29.

HOLLMANN SAMUEL MüLLER, K. E. F. H. N. Tabpfn: A transformer that solves small tabular classification problems in a second. 2023. Citado 2 vezes nas páginas 22 e 33.

LEMES, M. M.; LEMOS, A. N. L. E. O uso da inteligência artificial na saúde pela administração pública brasileira. *Cadernos Ibero-Americanos de Direito Sanitário*, v. 9, n. 3, p. 166–182, 2020. Citado na página 19.

LI, C. *A Gentle Introduction to Gradient Boosting*. Boston: [s.n.]. <http://www.chengli.io/tutorials/gradient_boosting.pdf>. Acesso em: 10 set 2023. Citado 2 vezes nas páginas 23 e 33.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. Citado na página 21.

NVIDIA. *XGBoost*. <https://www.nvidia.com/en-us/glossary/data-science/xgboost/>. Acesso em: 10 set 2023. Citado na página 25.

RIBEIRO, J. F. et al. Reestruturação das profissões da saúde e perspectivas para o futuro na era da inteligência artificial. *Comunicação em Ciências da Saúde*, Escola Superior de Ciências da Saúde, 2021. Citado na página 19.

ROFFO SIMONE MELZI, U. C. A. V. M. C. G. Infinite feature selection: A graph-based feature filtering approach. 2020. Citado 3 vezes nas páginas 18, 28 e 41.

SAUDE, M. da. *Acesso às informações do SIM (Sistemas de Informação de Mortalidade)*. Brasília: [s.n.]. <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10uf.def>. Acesso em: 10 set 2023. Citado na página 29.

SAUDE, M. da. *DataSUS*. Brasília: [s.n.]. <https://datasus.saude.gov.br/>. Acesso em: 10 set 2023. Citado na página 29.

SIDRA - Sistema IBGE de Recuperação Automática. Brasília: [s.n.]. <https://sidra.ibge.gov.br/tabela/7362/>. Acesso em: 10 nov 2023. Citado na página 45.

SOUZA, G. N. de; BULGARELI, J. V. Uso da inteligência artificial aplicada ao processo decisório na alocação de recursos na saúde pública do brasil: uma revisão integrativa da literatura. *JMPHC/ Journal of Management & Primary Health Care/ ISSN 2179-6750*, v. 15, n. spec, p. e012–e012, 2023. Citado na página 19.

TOUVRON, H. et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 45, n. 4, p. 5314–5321, 2023. Disponível em: <https://doi.org/10.1109/TPAMI.2022.3206148>. Citado 2 vezes nas páginas 26 e 39.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. [s.n.], 2017. p. 5998–6008. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>. Citado 2 vezes nas páginas 26 e 39.