Projeto de Formatura – 2023

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema:

Explorando Padrões de Mortalidade no Brasil, com Aprendizado de Máquina

Introdução

- As causas de morte refletem uma complexa interação sociais, biológicos e comportamentais
- A compreensão da relação entre esses fatores é fundamental para a construção de políticas de saúde pública e melhorar a qualidade e expectativa de vida da população
- Por meio dos dados públicos de saúde pesquisadores no Brasil e EUA já demonstraram a relação de fatores, como a cor da pele, nas causas de morte
- O aprendizado supervisionado é uma técnica que permite a partir de amostras categorizadas, descritas por diferentes atributos, prever, em um novo conjunto de dados com os mesmos atributos, qual a categoria das novas amostras
- Nesse trabalho, diferentes técnicas de aprendizado supervisionado foram aplicadas a amostras de falecimentos

Preliminares e Definição do Problema

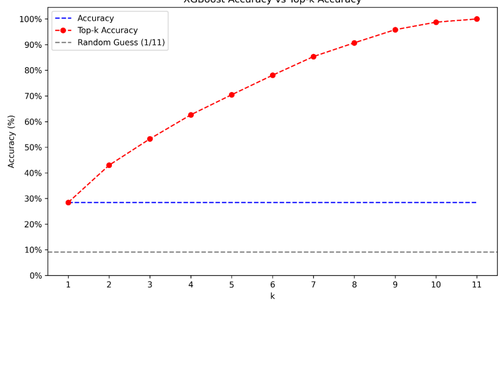
- No Brasil, os óbitos com diversos atributos são registrados nos Sistemas de Informação de Mortalidade (SIM), na plataforma DataSUS
- A Fiocruz, por meio da Plataforma de Ciência de Dados aplicada à saúde, extrai e enriquece as bases disponibilizadas pelo governo. A base final contém 159 atributos, dos quais muitos são categóricos, de todos os óbitos de 1996 a 2021
- Em virtude do número extenso de atributos e amostras, escolheu-se reduzir o espaço amostral para o Estado de SP
- Pelo número extenso de categorias (19), 9 das quais com menos de 1% de representatividade nas amostras, agrupou-se essas amostras em uma categoria “Outros”
- Pelo número ainda elevado de amostras, considerou-se a métrica de “top-k” como indicativo da qualidade do modelo
- A fim de reduzir o *bias* do modelo e reduzir o custo computacional do projeto, selecionou-se 1000 amostras de cada categoria para os dados de treinamento e teste

Questões de pesquisa:

1. Seria possível predizer a causa de morte de um falecido, a partir de um conjunto de dados que reflita particularidades do paciente e do óbito registrado?
2. Os modelos desenvolvidos generalizam os resultados obtidos para amostras externas aos casos de treinamento e teste?
3. Existem características que desempenham maior impacto na predição da causa da morte?

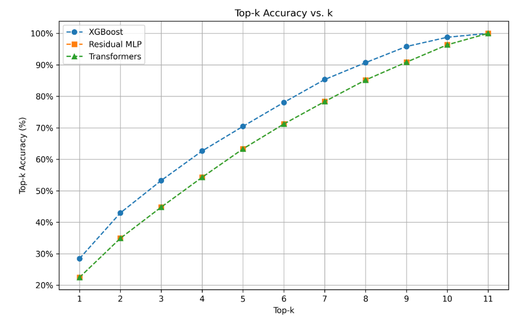
Experimentos (1/2)

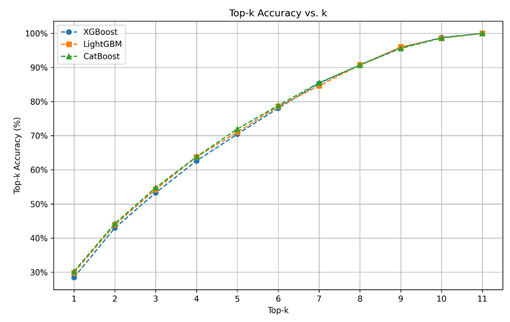
- Um modelo com XGBoosting é capaz de prever as categorias com uma precisão de 28,45%, ~3x acima do palpite aleatório (9,09%) e mais de 53,27% para o top-3



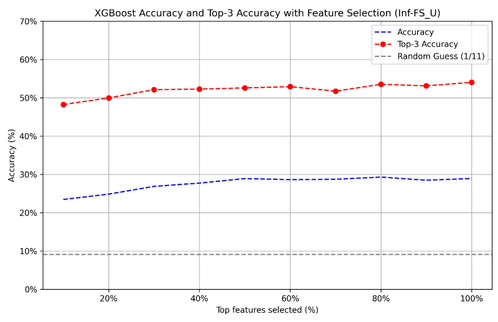
Experimentos (2/2)

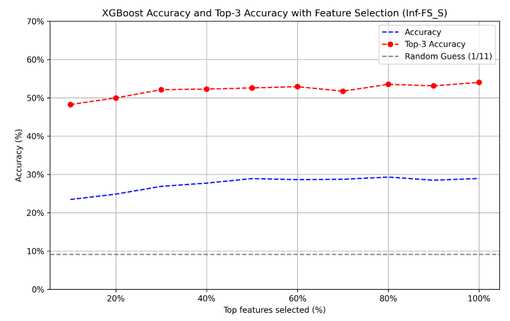
- Teste de diferentes modelos: modelos baseados na técnica de *boosting* performaram melhor



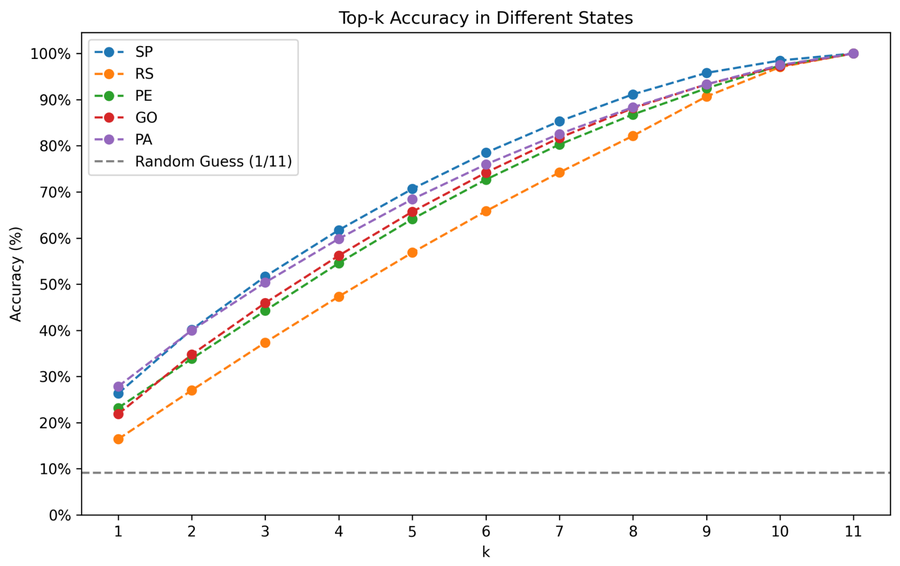


- Seleção de *features*: a precisão e a precisão top-3 varia muito pouco com o aumento da porcentagem das features selecionadas





- Teste em diferentes amostras: o modelo encontra uma precisão próxima do estado de São Paulo em 3 dos 4 estados avaliados, o que sugere que o modelo pode ser extrapolado a amostras fora do conjunto de treinamento e testes



Estado	Regular Accuracy	Top-3 Accuracy
Random guess	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%

Conclusões

- É possível predizer, a partir do conjunto de features selecionado, a causa de óbito de um falecido
- Os modelos desenvolvidos são capazes de generalizar os resultados obtidos para amostras aos casos de treinamento e teste, inclusive pode ser generalizado para outros estados da federação com características bem diferentes
- O modelo gerado possui resultado semelhante para apenas 10% das features e a precisão acima desse valor não cresce de maneira relevante

Integrantes: Filipe Penna Cerávolo Soares
Professor(a) Orientador(a): Prof. Dr. Artur Jordão