### Projeto Integrado de Aprendizado de Máquina

Pontifícia Universidade Católica de Minas Gerais

Professor: Geanderson Esteves dos Santos

#### Ponto de Controle 1

Marcelo Honório de Oliveira Mariana Mari Ikoma Sakamoto Nayara Mohana Rosa Ferreira

# 1 Introdução

O câncer do colo do útero é causado pela infecção persistente por tipos oncogênicos do Papilomavírus Humano (HPV). Na maioria dos casos, a infecção genital por HPV não evolui para lesões cancerosas, apesar de ocorrer frequentemente. Entretanto, os tipos 16 e 18 do vírus alteram mecanismos celulares que induzem a carcinogênese, evoluindo de lesão infecciosa para o câncer de colo uterino. Os fatores de risco estão associados à comportamentos como início precoce da atividade sexual, múltiplos parceiros sexuais, uso prolongado de pílulas anticoncepcionais e tabagismo. A detecção precoce do câncer é uma estratégia para diagnosticar o tumor na fase inicial e, assim, possibilitar maior sucesso no tratamento [1]. A proposta deste estudo é aplicar técnicas de aprendizado de máquina para extrair conhecimento a partir de um conjunto de dados para auxiliar no diagnóstico do câncer de colo uterino.

#### 2 Justificativa

O diagnóstico precoce do câncer de colo uterino depende de programas de saúde pública que realizem exames preventivos e de rastreamentos em mulheres portadoras dos comportamentos de risco. A detecção do tumor pode ser feita por meio da investigação com exames clínicos, diagnósticos laboratoriais ou por imagem. A biópsia é realizada em pacientes o qual o resultado do exame Papanicolau apresenta células anormais. Nosso estudo propõe modelos de aprendizagem de máquinas que prevejam o surgimento do câncer de colo uterino a partir de atributos clínico-patológico e comportamental. Assim, auxiliar na decisão médica otimizando o processo de diagnóstico e tratamento da doença.

# 3 Objetivos

Desenvolver e comparar modelos preditivos eficientes em prognosticar a biópsia do câncer de colo uterino.

### 4 Metodologia

O conjunto de dados [2] utilizado no estudo foi coletado no Hospital Universitário de Caracas em Caracas na Venezuela [3]. O mesmo inclui informações demográficas, hábitos e histórico médico de 858 pacientes. Após análise exploratória, será realizada a preparação dos dados, tratando os dados ausentes, aumentamdo os registros e balanceando as classes. A figura 1 representa o quadro de desbalanceamento da base original. Os métodos de classificação máquinas de vetores de suporte, arvore de decisão e modelagem de rede neural receberão ajustes, considerando a acurácia maior que 90% dos casos testados. O processo se dará de forma iterativa, o que pressupõe o retorno a etapas anteriores do processo para ajustes caso necessário.

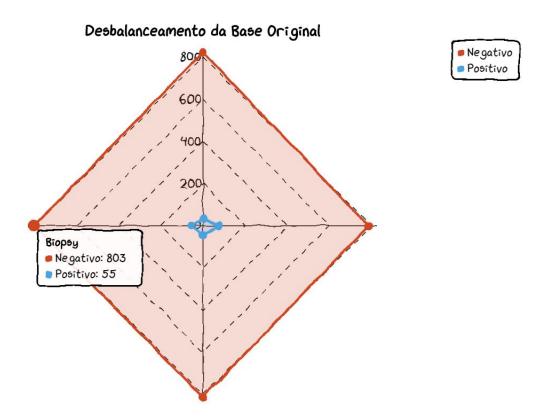


Fig. 1 - Classes desbalanceadas. Em vermelho, casos negativos para câncer de colo uterino. Em azul, casos positivos para câncer de colo uterino. À esquerda, resultado da biópsia. À direita, resultado do exame Schiller. Acima, o resultado do exame Hinselmann. Abaixo, resultado da citologia (Papanicolau).

#### 5 Referências

- 1. INCA https://www.inca.gov.br/assuntos/cancer-do-colo-do-utero (2020)
- 2. Cervical cancer (Risk Factors) Data Set https://archive.ics.uci.edu/ml/machine-learning-databases/00383/risk factors cervical cancer.csv
- 3. Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening." Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.