Seminário em Tecnologia da Informação do Programa de Capacitação Institucional (PCI) do CTI Renato Archer
* XIII Seminário PCI - Campinas, outubro de 2023 *

ANÁLISE E APLICAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NO BANCO DE DADOS PÚBLICO DO INSTITUTO NACIONAL DO CÂNCER (INCA)

Mariangela Dametto
Rodrigo Bonacin
Átila Kardec Alves (colaborador - CTI)
mdametto@cti.gov.br

INTRODUÇÃO

Milhares de casos de câncer são diagnosticados por ano no Brasil (BRASIL. Ministério da Saúde. DATASUS. Tabnet. Brasília, DF: Ministério da Saúde, 2022) e, segundo o Instituto Nacional do Câncer (INCA), estima-se que até 2025 serão mais de 700 mil casos anuais.

A literatura apresenta que diversos algoritmos de aprendizado de máquina têm sido utilizados com sucesso na classificação, predição e/ou clusterização com foco na análise de diagnóstico, prognóstico, detecção e classificação do câncer conforme as características fisiológicas dos pacientes (MANIKANDAN, P.; DURGA, U. & PONNURAJA, C., 2023; MOKOATLE, M., MARIVATE, V., MAPIYE, D. ET AL., 2023). Contudo, sabe-se que os diversos tipos de câncer podem se desenvolver devido à influência também de fatores genéticos, por isso nosso interesse em estudar principalmente os bancos de dados brasileiros que mantêm seus dados abertos e estão disponíveis na Internet.

OBJETIVO

Neste trabalho, aplicamos os algoritmos *Naive Bayes*, *Random Forest* e REPTTREE, visando encontrar um modelo robusto de classificação de cada tipo de câncer para as 7 classes que representam o estado da doença, utilizadas neste trabalho.

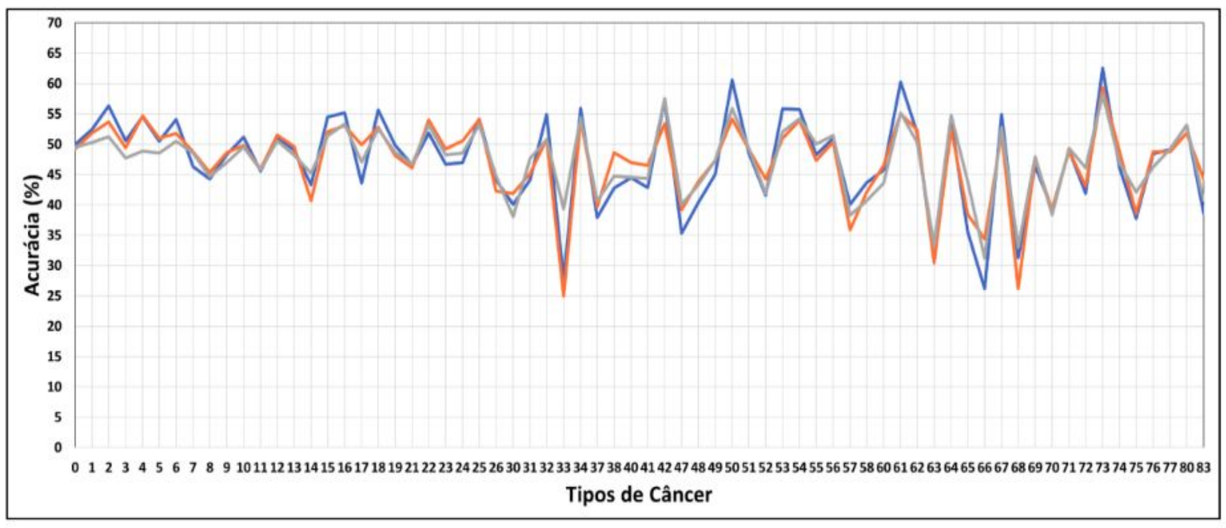
Futuramente, espera-se que possa haver uma junção do modelo treinado com a base brasileira aos modelos encontrados internacionalmente, havendo dessa forma, maior quantidade de dados que possam embasar e fundamentar a tomada de decisão por parte dos profissionais de saúde, seja para um diagnóstico mais preciso, para a escolha de um melhor tratamento especificamente para cada paciente, dentre outros propósitos.

MÉTODOS

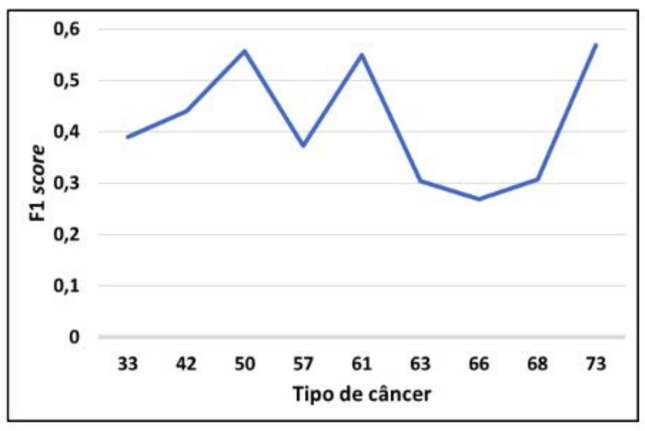
- importação do dicionário e do banco de dados do INCA (https://irhc.inca.gov.br/RHCNet/downloadTabWin!bases.action), novembro 2022.
- software POSTGRESQL (https://www.postgresql.org/) e aplicativo PGADMIN4 (https://www.pgadmin.org) usados para acesso à base de dados importada.
- análise, visualização e exploração das informações contidas nos atributos (*features*) existentes no banco de dados, seguida pelo pré-processamento dos dados.
- estabeleceu-se como classe o estado da doença, que continha sete classes: remissão completa da doença, remissão parcial do câncer, doença estável, progressão da doença, suporte terapêutico oncológico, não se aplica e sem informação.
- ao todo 90 tipos de câncer foram separados e estudados individualmente, e posteriormente eliminados os tipos com quantidade inferior a 200 registros.
- portanto, foram estudados 68 tipos de câncer com 2.888.861 registros de pacientes.

RESULTADOS

Na Figura abaixo estão mostrados valores de acurácia para cada um dos 68 tipos de câncer. A tendência para os 3 tipos de algoritmos são bastante semelhantes. E o câncer da glândula tireóide (CID 73 - 78.959 pacientes) é o que apresenta maior valor de acurácia (~60% - considerando as 7 classes), sendo o câncer de traquéia (CID 33 - 275 pacientes) o de menor valor de acurácia (~25% para o Naive Bayes – laranja, e Random Forest - cinza) e ~40% para o REPTTREE (azul).



Na Figura abaixo estão representados os valores de F1 *score* para 9 tipos de câncer separadamente, considerando as 7 classes. O algoritmo REPTTREE foi escolhido para esta comparação, pois este algoritmo apresentou os maiores e menores valores de acurácia.



CONCLUSÕES

O câncer de mama (CID 50), o de próstata (CID 61) e o da glândula tireóide (CID 73) apresentam maiores acurácias e também uma das maiores médias do F1 score considerando-se as 7 classes utilizadas.

Também para os CID 50, 61 e 73, a métrica F1 score está bastante balanceada entre as 7 classes. Isso sugere que o filtro de balanceamento das classes aplicado pode ter tido melhor desempenho para estes tipos de câncer, ou que estes já estariam com as classes mais balanceadas no próprio dataset, sendo que esta confirmação está sendo realizada atualmente.

REFERÊNCIAS

1. MANIKANDAN, P.; DURGA, U. & PONNURAJA, C. An integrative machine learning framework for classifying SEER breast cancer. Sci Rep 13, 5362, 2023.
2. MOKOATLE, M., MARIVATE, V., MAPIYE, D. ET AL. A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application. BMC Bioinformatics 24, 112, 2023.