XXXII Encontro Anual de Iniciação Científica
IX Encontro Anual de Iniciação Científica Jr.
EAIC 2023

ESTUDO DE FATORES DE RISCO ASSOCIADOS AO CÂNCER DE COLO CERVICAL POR MEIO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Igor Mateus Beledeli (CNPq), 22009726@uepg.br, Alaine Margarete Guimarães (Orientador), alainemg@uepg.br, Edneia Peres Machado, Andrea Timoteo dos Santos, Daniela Etlinger Colonelli

Universidade Estadual de Ponta Grossa/Departamento de Informática

Ciência da Computação-Metodologia e Técnicas da Computação.

Palavras-Chave: *câncer de colo cervical, exame Papanicolau, aprendizado de máquina, papilomavírus, mineração de dados.*

Resumo

O objetivo desse trabalho foi analisar, por meio de técnicas de *Machine Learning* (ML), uma base de dados de anamnese coletados de pacientes durante a realização do exame Papanicolau associados aos resultados das respectivas análises citopatológicas. Buscou-se identificar padrões de comportamento que podem contribuir na prevenção do câncer de colo cervical. O fato de haver muitas inconsistências na base de dados impactou nos resultados, confirmando que o sucesso ML é diretamente dependente da qualidade dos dados.

Introdução

O câncer de colo do útero, também chamado câncer cervical, é o quarto câncer mais comum em mulheres em todo o mundo (WHO, 2022). Entre os principais fatores que favorecem o desenvolvimento deste tipo de câncer estão a infecção pelo vírus Papilomavírus Humano (HPV). A prevenção e o diagnóstico precoce ainda são os melhores caminhos para vencer essa doença.

Este trabalho buscou, com a utilização de técnicas de aprendizado de máquina (ML – do inglês *Machine Learning*), reconhecer padrões em dados de anamnese de pacientes coletados durante a realização do exame Papanicolau associados aos resultados das respectivas análises citopatológicas. Buscou-se padrões de comportamento que podem contribuir para a prevenção do câncer de colo cervical.

Materiais e Métodos

Os dados utilizados são provenientes do Projeto de Extensão "Prevenção e Educação na Atenção à Saúde da Mulher: Coleta de Exame Papanicolau" (FARINA *et al.,* 2021), em execução na Universidade Estadual de Ponta Grossa, em parceria com instituições de saúde locais.

Devido ao fato de os dados terem sido coletados em seis locais diferentes, a base apresentou inconsistência, visto que algumas informações eram coletadas somente em alguns locais enquanto outras foram obtidas em locais diferentes, resultando em uma grande quantidade de dados faltantes e de campos categorizados sem legendas pré-definidas.

Foi realizado o pré-processamento no qual optou-se por dividir a tabela inicial em seis tabelas diferentes, de acordo com o local de coleta do exame, para minimizar a dificuldade de tratar tantos dados faltantes. Após isso, foram aplicados algoritmos de seleção de atributos por local, visando identificar os atributos mais relevantes. Em seguida, foi executado o algoritmo Apriori, para identificar associações entre os atributos selecionados, considerando as 10 regras mais relevantes para cada tabela. Por fim, foram aplicados algoritmos de classificação visado correlacionar os dados de anamnese como os resultados da análise citopatológica.

Para a análise dos dados e aplicação dos algoritmos de ML foi utilizada a plataforma Weka (FRANK *et al.*, 2016), incuindo a extensão AutoWeka (NANTASENAMAT *et al.* 2015).

Resultados e Discussão

A base de dados utilizada apresentou características bastantes distintas para cada subdivisão da base original. Os resultados do algoritmo de associação apresentaram um alto grau de confiança, variando de 0,91 à 1. Para a classificação, o melhor valor da estatística *Kappa* foi 0,8 e o menor erro, considerando a raiz quadrada do erro relativo, foi de 48,6%. Os melhores e piores resultados foram obtidos com os dados coletados no Ambulatório da UEPG e nas Unidades Básicas de Saúde (UBS), respectivamente.

Conclusão

A grande quantidade de dados faltantes na tabela fornecida foi fator limitante para a obtenção de melhores resultados para a classificação, o que confirma que a qualidade do aprendizado de máquina depende da qualidade dos dados.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Iniciação Científica concedida.

Referências

WHO, World Health Organization. **Cervical Cancer**. 22 feb. 2022. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer. Acesso em: 11 set. 2023.

FARINA, D. S. **Ficha de anamnese nas consultas médica e de enfermagem na prevenção do câncer de colo do útero em projeto extensionista.** Trabalho de Conclusão de Curso (Graduação em Enfermagem) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2021.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench**: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". 4. ed. Morgan Kaufmann, 2016.

NANTASENAMAT, C. *et al*. AutoWeka: Towards an Automated Data Mining Software for QSAR and QSPR Studies. *In*: Cartwright H. **Artificial Neural Networks**: Methods and Applications. 2. ed. Oxford: Oxford University, 2015. v. 1260, p. 119-147.