UNIVERSIDADE FEEVALE

FERNANDO BRUNNER

Aplicação do aprendizado de máquina na saúde. Desenvolvimento de um modelo preditivo para o diagnóstico do câncer de mama

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo

2021

FERNANDO BRUNNER

Aplicação do aprendizado de máquina na saúde. Desenvolvimento de um modelo preditivo para o diagnóstico do câncer de mama

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial

à obtenção do grau de Bacharel em

Ciência da Computação pela

Universidade Feevale

Orientador: Ricardo Ferreira de Oliveira

Novo Hamburgo

2021

RESUMO

 O câncer de mama, hoje, é um dos tipos de cânceres mais comuns que existem. Este câncer mata todos os anos milhares de pessoas. Este tipo de câncer acomete principalmente mulheres. Se descoberto cedo, o paciente tem grandes chances de sobreviver à doença. Porém, nos dias atuais, mesmo com o avanço tecnológico, muitos casos ainda têm o diagnóstico tardio, aumentando assim, a mortalidade da doença. Falhas humanas, com profissionais mal preparados e falta de infraestrutura são algumas das causas que dificultam, principalmente para populações mais pobres, o diagnóstico da doença. Nos últimos anos, o ramo do aprendizado de máquina ganhou muita popularidade na aplicação à saúde, por ser uma solução viável para esses problemas. Capaz de analisar grandes volumes de dados, fazer predições e tomar decisões com uma baixa taxa de erro, com uso de aprendizagem supervisiona, *computer vision* e redes neurais, *machine learning* é usado para detectar e diagnosticar diversos tipos de cânceres, incluindo o câncer de mama. O objetivo deste trabalho é criar um modelo preditivo para o diagnóstico de tumores de mama, utilizando *python* e o *scikit-learn* para o desenvolvimento e para o treinamento do modelo utilizar uma base com dados retirados de lâminas de exames de tumores. O modelo utilizará técnicas de aprendizado de máquina e aprendizado supervisionado.

Palavras-chave: Inteligência Artificial. *Machine Learning*. Aprendizado Supervisionado.

SUMÁRIO

MOTIVAÇÃO ...........................................................................................................................5

OBJETIVOS ..............................................................................................................................8

METODOLOGIA ......................................................................................................................9

CRONOGRAMA ....................................................................................................................10

BIBLIOGRAFIA ....................................................................................................................11

 MOTIVAÇÃO

 Segundo o instituto nacional de câncer, o câncer de mama é uma doença causada pela multiplicação desordenada de células da mama. Este crescimento desordenado causa a multiplicação de células anormais. É um dos mais comuns e perigosos tipos de câncer, matou 18.295 pessoas no ano de 2019, somente no Brasil (2019 - Atlas de Mortalidade por Câncer), e atingiu 66.280 novos casos entre mulheres no ano de 2020 no Brasil (INCA - Instituto Nacional do Câncer). Este número representa, aproximadamente, 44 casos para cada 100.000 mulheres. É o quinto tipo de câncer que mais mata entre a população geral e o mais mortal entre a população feminina. Em todas as regiões brasileiras o câncer de mama feminino já é o mais comum, com ocorrência de 81.06 por 100 mil mulheres na Região Sudeste, de 71.16 por 100 mil mulheres na Região Sul, de 45.24 por 100 mil mulheres na Região Centro-Oeste, de 44.29 por 100 mil mulheres na Região Nordeste, e de 21.34 por 100 mil mulheres na Região Norte (Estimativa | 2020. Incidência do Câncer no Brasil. Instituto Nacional de Câncer José Alencar Gomes da Silva: INCA, 2020).

 Descoberto em estágios iniciais, as chances de sucesso no tratamento do câncer de mama são grandes. Porém, o diagnóstico tardio ainda ocorre em uma parcela significativa dos casos, o que impossibilitaa recuperação da doença. Entre as causas para a dificuldade de diagnóstico, estão a falta de profissionais com capacitação necessária e a falta de infraestrutura para suprir a demanda. O tumor pode ser classificado entre benigno ou maligno. Os tumores classificados como benignos não são considerados cancerígenos, diferentemente dos tumores malignos. Caso o tumor maligno não seja identificado precocemente e tratado, ele pode se espalhar para outras partes do corpo. O câncer de mama é classificado como um tumor maligno que se desenvolve nas células da mama. Para a classificação de tumores malignos e benignos são levados em consideração a aparência, a estrutura e o comportamento reprodutivo.

 Com o avanço da tecnologia, soluções para os problemas anteriormente apresentados surgiram. A inteligência artificial, descrita porJohn McCarthy(1956) como "Ciência e engenharia de fazer máquinas inteligentes" (tradução livre), ajuda na área da saúde a melhorar a precisão e a velocidade de diagnósticos e a prevenir contra falha humana e também ajudando no prognóstico médico. Com a ascensão de campos da inteligência artificial, como *Machine Learning* e *Deep Learning*, grandes volumes de dados podem ser analisados e algoritmos complexos podem ser usados para prática médica.

 Como dito anteriormente, *machine learning* é um tipo de Inteligência Artificial é um campo da ciência da computação, que permite que desenvolvedores escrevam código de uma maneira mais simples. De forma simplista, *Machine Learning* foca em ensinar a máquina a mudar e aprender sempre que é exposta a novos dados. Arthur Samuel, em 1959, definiu o aprendizado de máquina como sendo "O campo de estudo que dá a computadores a habilidade de aprender sem ser explicitamente programados para isso".(Tradução livre).

*Machine Learning* explora algoritmos que podem aprender e aperfeiçoar seus resultados e fazer previsões sobre grandes volumes de dados. "Um programa aprende pela experiência E, a respeito de algumas classes de tarefas T, e performance medida por P, se o seu desempenho nas tarefas T, como medido em P, melhoram com a experiência E" (MITCHELL T. M. Machine Learning. 1997)(Tradução livre).

 Existem três principais tipos de aprendizado de máquina: Aprendizado supervisionado, Aprendizado não supervisionado e aprendizado por reforço. O aprendizado supervisionado acontece quando passamos para o modelo os resultados já definidos, para que ele possa aprender a partir desses dados pré-definidos quais devem ser os resultados de saída esperados. Esses dados pré-definidos servem de "supervisão" para o modelo possibilitando ajustes na previsão. Como exemplos de aprendizado supervisionado são regressão linear e redes neurais artificiais. Ao contrário do modelo de aprendizado supervisionado, o modelo de aprendizado

não supervisionado, não utiliza resultados pré-definidos durante o processo de aprendizado. O modelo tentará encontrar por conta própria o melhor resultado. Como exemplo de aprendizado

não supervisionado temos o *autoencoders*, deep *belief nets*. O aprendizado por reforço tenta maximizar o resultado em uma situação específica. Baseado no *Markov decision processes*, no aprendizado por reforço a decisão será tomada de acordo com cada situação, definidos por uma combinação de ações e pelo ambiente.

 De acordo com o trabalho de Hellen Geremias dos Santos et al. (2019) que estudou a aplicação de algoritmos de aprendizado de máquina aplicados a saúde e desenvolveu um modelo preditivo utilizando *Machine Learning* em um cenário supervisionado para a predição de óbito de idosos no estado de São Paulo, levando em consideração fatores de risco na velhice, e alcançou resultado satisfatório no desempenho de todos os algoritmos utilizados, temos uma ideia de como utilizar algoritmos de aprendizado de máquina com aprendizado supervisionado para a construção do modelo preditivo para alcançar o objetivo deste trabalho.

 Tendo em vista as questões anteriores, o objetivo do trabalho será desenvolver um modelo preditivo em *python*, utilizando *jupyter notebook* e um *dataset* com dados sobre tipos de tumores para a classificação de tumores de mama femininos. Também há o interesse do autor na oportunidade de aprender e aprofundar o conhecimento na área do aprendizado de máquina, ciência de dados e aperfeiçoar os conhecimentos em *python* e em todas as ferramentas envolvidas no desenvolvimento do trabalho.

OBJETIVOS

Objetivo geral

 O objetivo deste trabalho é criar um modelo preditivo para o diagnóstico de tumores de mama, utilizando *python* e o *scikit-learn* para o desenvolvimento e para o treinamento do modelo utilizar uma base com dados retirados de lâminas de exames de tumores. O modelo utilizará técnicas de aprendizado de máquina e aprendizado supervisionado.

Objetivos específicos

* Realizar pesquisa bibliográfica sobre *Machine Learning*
* Realizar pesquisa bibliográfica sobre *Machine Learning* aplicado à saúde
* Desenvolver o modelo preditivo em *python*
* Treinar o modelo utilizando a base de dados
* Avaliar os resultados utilizando outros trabalhos na área
* Validação do modelo preditivo

METODOLOGIA

Baseando-se em Prodanov e Freitas (2013) que descreveram os conceitos de metodologia científica, o trabalho desenvolvido tem por característica natureza de pesquisa aplicada, levando em consideração que será desenvolvido um modelo preditivo em *python*, treinado em cima de dados de um *dataset*, para o diagnóstico de câncer de mama. Sendo pesquisa aplicada definida como "objetivo de gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais" (PRODANOV E FREITAS, 2013)

Os objetivos classificam o trabalho como pesquisa exploratória, que é definida como "quando a pesquisa se encontra na fase preliminar, tem como finalidade proporcionar mais informações sobre o assunto que vamos investigar, possibilitando sua definição e seu delineamento"(PRODANOV E FREITAS, 2013) , já que envolve o estudo para maior compreensão do tema do trabalho para construção do modelo preditivo.

Para fundamentar os conhecimentos teóricos e servir de base para a construção do modelo preditivo, será realizada a pesquisa bibliográfica. Também serão feitas pesquisas de aprendizado de máquina e aprendizado supervisionado para o serem usados no modelo e por fim, serão usados trabalhos correlatos para validar os resultados e avaliar possíveis melhorias no desempenho do modelo desenvolvido.

CRONOGRAMA

Trabalho de Conclusão I

|  |  |
| --- | --- |
| Etapa  | Meses |
| Mar | Abr | Mai | Jun |
| Busca da base de dados | X |  |  |  |
| Leitura de trabalhos correlatos | X | X |  |  |
| Pesquisa Bibliográfica | X | X | X | X |
| Pesquisa de modelos preditivos |  | X | X | X |
| Escrita do TCC I |  | X | X | X |
| Revisão do TCC I |  |  | X | X |
| Entrega do TCC I |  |  |  | X |

Trabalho de Conclusão II

|  |  |
| --- | --- |
| Etapa  | Meses |
| Ago | Set | Out | Nov |
| Desenvolvimento do modelo | X | X | X |  |
| Treinamento do modelo |  | X | X |  |
| Avaliação do modelo |  |  | X | X |
| Escrita do TCC II | X | X | X | X |
| Revisão do TCC II | X | X | X | X |
| Entrega do TCC II |  |  |  | X |

BIBLIOGRAFIA

Câncer de mama. Instituto Nacional do Câncer, 2021. Disponível em: <"https://www.inca.gov.br/assuntos/cancer-de-mama">. Acesso em: 26 mar 2021

Dos Santos. H. G. et al. (2019) ***Machine learning* para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil**. Disponível em: <<https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2019000904002>> Acesso em: 27 mar 2021

INSTITUTO NACIONAL DE C NCER JOSÈ ALENCAR GOMES DA SILVA. **Estimativa** 2020. Disponível em: <<https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2020-incidencia-de-cancer-no-brasil.pdf>>

John McCarthy In: WIKIPÉDIA: a enciclopédia livre. Wikimedia, 2021. Disponível em: <https://en.wikipedia.org/wiki/John\_McCarthy\_(computer\_scientist)>. Acesso em: 25 mar. 2021.

MITCHELL, T. M. Título: **Machine Learning** Local de publicação: McGraw-Hill Science/Engineering/Math Março 1, 1997

PRODANOV, C. C.; FREITAS, E. C. De. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. Disponível em: <https://www.feevale.br/Comum/midias/0163c988-1f5d-496f-b118-a6e009a7a2f9/E-book Metodologia do Trabalho Cientifico.pdf>

SAMUEL, A. L. **Some Studies in Machine Learning Using the Game of Checkers** 1959. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560>