APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA IDENTIFICAÇÃO DE TUMORES MALIGNOS E BENIGNOS

Larissa Freitas da Silva
Unimar – Universidade de Marília
b.larissafs@gmail.com

RESUMO
A tecnologia evolui cada vez mais, com isto uma vasta gama de ramificações e ferramentas para que a sociedade usufrua de modo com que facilite no cotidiano, com criação de rotinas e automatizações. Por exemplo, os algoritmos de aprendizagem de máquina são promissores em vários ramos, o artigo apresenta uso da aprendizagem de máquina para analisar dados com desafio distinguir tumores malignos e benignos para melhorar os diagnósticos usando classificadores de aprendizagem de máquina.

Palavras-Chave: Tecnologia, Aprendizagem de máquina, Automatizações. Análise.

1. INTRODUÇÃO

Mais de 70% de casos de câncer de mama no Brasil, são diagnosticados em estágio avançado com possivel ligação no desenvolvimento socioeconômico e tecnológico. De acordo com Camila Veras Mota, a tecnologia pode ser eficaz para diagnósticos de câncer de mama, trazendo mais qualidade de vida e saúde, contribuindo com os diagnósticos. Como o fator tecnológico poderia ajudar com a análise de dados e aprendizado de máquina tornando mais eficaz o diagnóstico do câncer de mama?
De acordo com a revista científica Nature, Fergus Walsh obteve a seguinte informação para a BBC em 2 de janeiro de 2020: “A inteligência artificial é mais precisa do que os médicos de câncer de mama a partir de uma mamografia”, por meio de um estudo em que uma universidade de Londres treinou um modelo com dados de 29 mil mulheres superando seis radiologistas com a leitura das mamografias.
Diante deste problema, o artigo trás uma breve abordagem para clareza do uso dos dados utilizados, ferramentas, modelos e técnicas que poderiam ser usados como base de pesquisa e análise de tumores e assim mostrar a eficiência que pode trazer para os diagnósticos de câncer de mama benigno e maligno.

2. METODOLOGIA

O foco é interagir com a aprendizagem de máquina, buscando desenvolver técnicas para a leitura dos diagnósticos, como o uso das ferramentas Visual Studio Code e bibliotecas com desempenho e performance para ciência de dados, leitura, análise, manipulação e filtragem de

dados, são elas Numpy, Pandas, além da biblioteca Sckit-learn para aprendizado de máquina com funções para trabalhar com árvores de decisão e treinamento com diversos outros modelos, todas essas bibliotecas mantém bem documentadas todas as funções descritas para melhor usabilidade.

Para a metodologia num mapeamento dos diagnósticos para poder ser selecionadas métricas que farão simulações no treinamento com vários conjuntos, há o uso dos dados do dataset breast-cancer, dataset de câncer de mama com diagnósticos, sendo eles positivos e negativos, concluindo a análise de classificação desses tumores usando aprendizado de máquina, com 3 classificadores, sendo eles Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Naive Bayes (NB).

Para determinar se os resultados dos classificadores foram bons e fazer uma comparação, foi preciso avaliar algumas métricas, para posteriormente medir o desempenho na matriz de confusão para analisar a possibilidade de o classificador confundir um paciente saudável com um paciente doente ou vise-versa, assim determinando de modo analítico sua acurácia.

O uso do matplotlib seria apenas um exemplo de como visualizar os dados de forma mais intuitiva, o matplotlib é uma biblioteca que possibilita a criação de visualizações estáticas, sendo gráficos, figuras interativas, entre outras possibilidades, que auxiliam na visualização de informações. Mais sobre o matplotlib pode ser encontrado em sua documentação.

4. CONCLUSÃO

Para comparação de resultados, pode ser pelas métricas de resultado ou por exemplo com base na leitura da matriz de confusão, logo, com este pequeno exemplo de como pode ser feita a leitura na matriz de confusão:

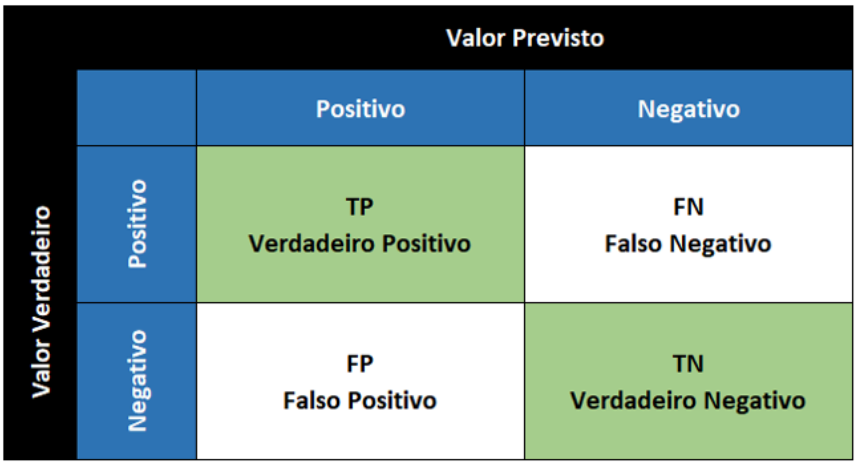

Imagem extraída do artigo Machine Learning pt.3: Descomplicando a Matriz de confusão e Calculando Acurácia e Recall de Leonardo Karpinski no Linkedin.

Com isso percebe-se que Naive Bayes se adequou melhor aos dados, ao considerar as métricas, com média na *precision* macro de 0.94, um *recall* médio de 0.92, *F1* médio de 0.93 e Acurácia média de de 0.93, além de considerar todas as saídas da *confusion matrix*, que mostra que o classificador Naive Bayes obteve melhor desempenho, mostrando menos falso-positivos e falso-negativos, e melhores resultados em verdadeiro-positivos e verdadeiro-negativos.

REFERÊNCIAS

MOTA, Camila Veras – Por que mais de 70% dos casos de câncer de mama no Brasil são diagnosticados em estágio avançado. BBC News Brasil, 2019. Disponível em: https://www.bbc.com/portuguese/brasil-49966596. Acesso em: 28 abr 2024.

WALSH, Fergus – Câncer de mama: Inteligência artificial bate médicos em diagnósticos. BBC, 2020. Disponível em: https://www.bbc.com/portuguese/geral-50971176. Acesso em: 28 abr 2024.

M Yasser H, Breast Cancer Dataset. Kaggle. 2020. Disponível em: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.

KAPINSKI, Leonardo - Machine Learning pt 3: Descomplicando a Matriz de Confusão e Calculando Acurácia, Precisão e Recall. Linkedin, 2018. Disponível em: https://pt.linkedin.com/pulse/machine-learning-pt-3-descomplicando-matriz-de-e-recall-karpinski