Estudo de Técnicas de Inteligência Artificial Aplicadas à Análise de Dados de Saúde e Reconhecimento de Faces

Guilherme Webster Chamoun[1,2], **Guilherme Cesar Soares Ruppert**[1]

[1]Divisao de Metodologias da Computação – DIMEC
CTI Renato Archer – Campinas/SP

[2]Universidade Estadual de Campinas - Campinas/SP

g257111@dac.unicamp.br,guilherme.ruppert@cti.gov.br

Abstract. *This article presents a brief study of machine learning techniques and tools, focusing on health data analysis. Using a public healthcare dataset available on the Kaggle platform, we present supervised classification experiments. In addition, an implementation was also developed for the problem of face recognition in images."*

Resumo. *Este artigo apresenta um breve estudo de técnicas e ferramentas de aprendizado de máquina com o foco análise de dados de saúde. Utilizado um conjunto de dados públicos de assistência médica disponíveis na plataforma Kaggle, apresentamos experimentos realizados de classificação supervisionada. Além disso, também foi desenvolvido uma implementação para o problema de reconhecimento de faces em imagens.*

1. Introdução

Inteligência artificial (IA) tem ganhado relevância nas mais diversas áreas. O termo refere-se a métodos computacionais para simular a capacidade humana de aprender, perceber e tomar decisões. Tais métodos variam dos mais simples, como uma regressão linear, até os mais sofisticados, como redes neurais.

Na área da saúde, a IA pode auxiliar na análise de grandes volumes de dados clínicos, detecção precoce de doenças, suporte à decisão médica, otimização de recursos hospitalares e descoberta de padrões que seriam difíceis de identificar manualmente. Essas aplicações permitem desde a triagem automatizada de exames de imagem até a predição de risco de pacientes desenvolverem certas condições, aumentando a eficiência e potencialmente salvando vidas.

Os métodos de aprendizado podem ser supervisionado e não supervisionado, sendo o supervisionado aquele que utiliza dados rotulados e o não supervisionado o que utiliza dados não rotulados.

Neste trabalho utilizamos apenas o aprendizado supervisionado, apresentando um estudo de análise de dados de saúde e também técnicas de reconhecimento de faces. Na Seção 2, descrevemos as ferramentas e técnicas utilizadas, incluindo bibliotecas e algoritmos de aprendizado de máquina. Na Seção 3, apresentamos os resultados obtidos com dados tabulares e imagens faciais. Por fim, na Seção 4, discutimos as conclusões e perspectivas para trabalhos futuros.

2. Ferramentas e Técnicas

Entre as ferramentas utilizadas, podemos citar o Google Colab, onde foram feitos os Jupyter Notebooks das atividades, a matplotlib que é uma biblioteca do Python para fazer gráficos e figuras, scikit-learn que é uma biblioteca para machine learning especificamente, e pandas que é uma biblioteca do python para manipular dados.

Scikit-learn: O `scikit-learn` é uma biblioteca em Python amplamente utilizada para aprendizado de máquina. Ela oferece implementações eficientes de algoritmos de classificação, regressão, clusterização e redução de dimensionalidade, além de ferramentas para pré-processamento de dados, avaliação de modelos e ajuste de hiperparâmetros. Sua interface consistente facilita a experimentação e comparação entre diferentes modelos.

Pandas: O `pandas` é uma biblioteca em Python para manipulação e análise de dados. Ela fornece estruturas como o `DataFrame`, que permite armazenar e operar sobre dados tabulares de forma intuitiva, além de funções para filtragem, agregação, tratamento de valores ausentes e integração com outras bibliotecas de análise e visualização.

Métrica ROC: A Curva ROC (*Receiver Operating Characteristic*) é uma ferramenta para avaliar a performance de classificadores binários. Ela plota a taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos (1-especificidade) para diferentes limiares de decisão. A área sob a curva (AUC) resume a qualidade do modelo: quanto mais próxima de 1, melhor o desempenho. Vale notar que na área da saúde os dados são sempre desbalanceados (ou seja, temos um maior número de pessoas consideradas normais para um problema de saúde específico), assim o uso de métricas alternativas a acurácia, por exemplo, é muito importante.

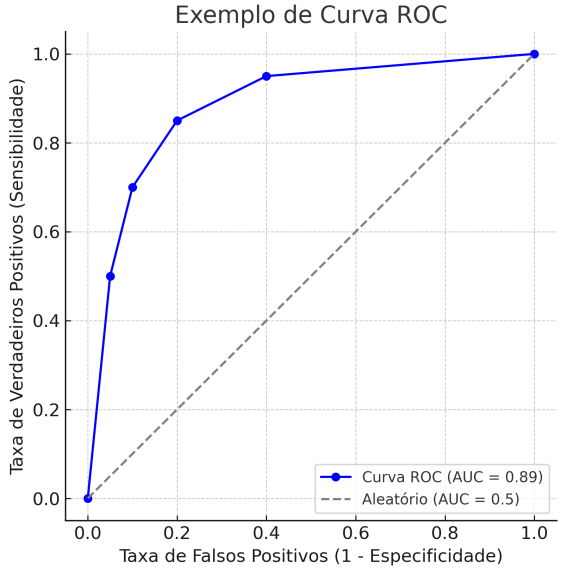

Figura 1. Exemplo de curva ROC com diferentes limiares de decisão.

Foram utilizadas muitas técnicas associadas a área de machine learning, como one-hot encoding para variáveis categóricas, normalização para variáveis numéricas, balanceamento dos resultados para o conjunto de treinamento, uso do algoritmo de Boruta

para seleção de variáveis, comparação entre diferentes classificadores como Redes Neurais, XGBoost, e Random Forest, GridSearch para ajuste de hiperparâmetros, e uso de valores SHAP para interpretabilidade do modelo.

Neste trabalho, foram estudados os seguintes classificadores e técnicas importantes para a ciência de dados:

- **Redes Neurais**: Redes neurais artificiais são modelos computacionais inspirados no funcionamento do cérebro humano, formadas por unidades chamadas neurônios artificiais, organizadas em camadas interconectadas. Elas recebem dados de entrada, processam-nos por meio de cálculos matemáticos e funções de ativação, e produzem uma saída ajustada de acordo com o problema em estudo. O processo de aprendizado ocorre pelo ajuste dos pesos dessas conexões, permitindo que a rede melhore seu desempenho ao longo do tempo. Graças a essa capacidade de aprender padrões complexos, redes neurais são amplamente utilizadas em áreas como reconhecimento de imagens e voz, análise de dados de saúde, processamento de linguagem natural e sistemas de recomendação.
- **Random Forest**: O Random Forest é um algoritmo de aprendizado de máquina baseado em um conjunto de árvores de decisão, utilizado principalmente para tarefas de classificação e regressão. Ele funciona criando várias árvores independentes a partir de amostras diferentes dos dados e, em seguida, combina os resultados para gerar uma previsão mais robusta e precisa. Essa abordagem reduz o risco de overfitting, comum em árvores de decisão individuais, pois a diversidade entre as árvores melhora a capacidade de generalização do modelo. Devido à sua eficiência e facilidade de interpretação, o Random Forest é amplamente aplicado em áreas como análise de dados médicos, detecção de fraudes, previsão de mercado e reconhecimento de padrões.
- **XGBoost**: Implementação otimizada do algoritmo *gradient boosting*, que é um método de aprendizado de máquina que combina vários modelos fracos, geralmente árvores de decisão pouco profundas, para formar um modelo forte e de alta precisão. Diferente do Random Forest, que constrói árvores em paralelo e depois agrega os resultados, o Gradient Boosting constrói as árvores de forma sequencial, onde cada nova árvore é treinada para corrigir os erros cometidos pelas anteriores. Esse processo é guiado pelo cálculo do gradiente de uma função de perda, o que permite ao modelo melhorar gradualmente o desempenho a cada iteração.
- **GridSearch**: Técnica sistemática para encontrar a melhor combinação de hiperparâmetros, treinando o modelo para cada combinação possível em um conjunto definido e avaliando seu desempenho.
- **SHAP**: Os valores de Shapley (SHAP) são um conceito emprestado da teoria dos jogos. Originalmente, os valores de Shapley eram usados para atribuir de forma justa a contribuição de um jogador ao resultado final de um jogo.

Ademais, foram importantes as noções de SVM's (Support Vector Machine), seu uso no reconhecimento de autofaces, e a importância de levar em conta diferentes métricas de performance como acurácia, precisão, recall e f1-score.

3. Resultados:

3.1. Análise de Dados Tabulares

Para este experimento, utilizamos um conjunto de dados educacional sintético, criado para simular registros de pacientes e disponível publicamente na plataforma Kaggle[1]. O dataset contém 10.000 amostras e 15 atributos, incluindo informações demográficas (e.g., Age, Gender), clínicas (e.g., Blood Type, Medical Condition) e administrativas (e.g., Hospital, Insurance Provider, Billing Amount). O objetivo do problema é prever a variável alvo Test Results, que pode assumir um de três valores: 'Normal', 'Abnormal' ou 'Inconclusive'.

O pré-processamento dos dados foi uma etapa crucial. Variáveis categóricas nominais, como Gender, Blood Type, Medical Condition, Admission Type, Insurance Provider e Medication, foram transformadas em formato numérico utilizando o OneHotEncoder da biblioteca scikit-learn. Este método cria uma nova coluna binária para cada categoria única da variável original, evitando a introdução de uma relação ordinal artificial que poderia ser mal interpretada pelos algoritmos.

O conjunto de dados completo foi então dividido em 70% para treinamento e 30% para teste. É preciso garantir um balanceamento nas classes da variável alvo para evitar viés no modelo, mas elas já estavam balanceadas. Usando python e scikit-learn, foram implementados experimentos para avaliar três classificadores (Gradient Boosting, Random Forest e uma Rede Neural simples).

Tabela 1. Desempenho Comparativo dos Classificadores (AUC ROC Média)

Modelo	AUC ROC (Média)
XGBoost	0.50
Random Forest	0.53
Rede Neural	0.50

Os resultados inciais deste experimento indicaram que os modelos baseados em árvores de decisão superaram a Rede Neural neste problema, com a Random Forest alcançando o melhor desempenho.

Com base nesses resultados, o modelo Random Forest foi selecionado para a etapa de otimização de hiperparâmetros via GridSearch. Após uma busca exaustiva por combinações ótimas de parâmetros como a a profundidade máxima das árvores, o modelo otimizado alcançou uma performance igual a pré-otimização.

3.2. Reconhecimento de Faces

Para o segundo experimento, foi abordada uma tarefa de reconhecimento de faces utilizando o conjunto de dados "Labeled Faces in the Wild"(LFW), que é disponibilizado diretamente pela biblioteca scikit-learn através da função fetch_lfw_people. O LFW é uma coleção de imagens de figuras públicas. Para criar uma tarefa de classificação tratável e com múltiplas amostras por classe, o dataset foi pré-processado com o parâmetro min_faces_per_person=70, uma prática padrão que filtra o conjunto para reter apenas indivíduos que possuem pelo menos 70 imagens distintas.

[1]https://www.kaggle.com/datasets/prasad22/healthcare-dataset?resource=download

O subconjunto resultante foi dividido em 75% para treinamento e 25% para teste, utilizando estratificação para garantir que todos os indivíduos estivessem proporcionalmente representados em ambos os conjuntos. Um modelo de Máquina de Vetores de Suporte (Support Vector Machine - SVM) com um kernel não-linear (”rbf”) foi treinado para classificar as imagens. Após o treinamento, o modelo foi avaliado no conjunto de teste e alcançou uma acurácia de 88%.

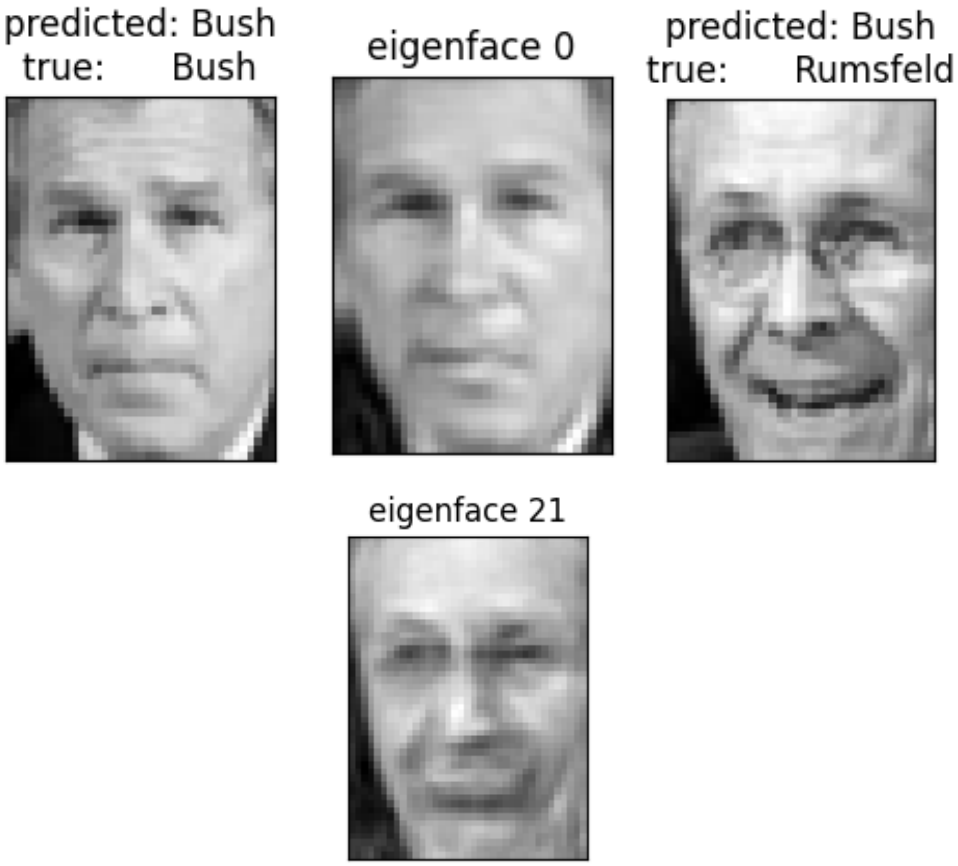

Figura 2. Exemplos de predições com SVM: acerto (esq.), eigenface-acerto (centro) e erro (dir.)., eigenface-erro (abaixo)

4. Conclusão

Podemos notar como o desempenho de algoritmos de machine learning é importante para diferentes áreas, ficando clara sua colaboração tanto em matérias de saúde quanto em algoritmos de reconhecimento de imagens.

Vale ressaltar que este é um estudo com escopo introdutório e existem métodos mais sofisticados que poderiam potencialmente alcançar resultados muito superiores. Ademais, é crucial reiterar que, no contexto da área da saúde, as ferramentas computacionais de apredizado de máquina podem contribuir muito os médicos no diagnóstico, tratamento ou conduta clínica.

Referências

- `https://www.coursera.org/learn/machine-learning/home/`

- https://www.coursera.org/learn/advanced-learning-algorithms/home/
- https://www.coursera.org/learn/unsupervised-learning-recommenders-home/
- https://www.coursera.org/specializations/deep-learning-healthcare
- https://www.youtube.com/playlist?list=PLeFetwYAi-F9Z0N3ZSjPL2i24PO5iiXaY
- https://www.youtube.com/playlist?list=PLpvV74h3lihLdYrlnhlx_phy4pFZeZsKx
- https://youtu.be/wB9C0Mz9gSo?si=dssZPuf30jxQbWTZ