Comparative Analysis of Machine Learning Algorithms in Predicting Smoking and Drinking Behaviors

Vitor Albuquerque de Paula
Programa de Pós-Graduação em Sistemas Inteligentes
Universidade Federal de São Paulo (UNIFESP)
São Paulo, Brasil
vitor.ap@proton.pm

***Abstract*—This study explores the application of various machine learning algorithms in predicting smoking and drinking behaviors based on health data. Employing a comparative analysis, the research assesses the performance of models like K-Nearest Neighbors, Random Forest, and Multi-Layer Perceptron. The results highlight the effectiveness of these algorithms in interpreting complex health data, with particular focus on the superior performance of MLPClassifier in capturing intricate data relationships. This work contributes to understanding how different machine learning approaches can be strategically utilized in health behavior predictions, offering insights into their applicability in medical informatics.**

***Index Terms*—Machine Learning, Health Data, Smoking Behavior, Drinking Behavior, MLPClassifier, Random Forest, K-Nearest Neighbors, Predictive Analysis**

I. Introdução

II. Objetivo do Projeto

O objetivo central deste projeto é desenvolver e aprimorar modelos de aprendizado de máquina que possuam a capacidade de prever comportamentos de risco, especificamente relacionados ao consumo de tabaco e álcool. Utilizamos para isso um robusto conjunto de dados oriundos de inquéritos de saúde pública, enriquecidos por variáveis biométricas detalhadas. A precisão na previsão desses comportamentos tem o potencial de transformar significativamente as estratégias de intervenção e prevenção em saúde pública.

III. Contexto e Importância

O consumo de tabaco e álcool são amplamente reconhecidos como fatores de risco para uma série de doenças crônicas e condições de saúde adversas. A modelagem eficaz de dados relacionados a esses comportamentos é uma ferramenta inestimável para a prevenção e o planejamento de intervenções em saúde pública. Neste contexto, a emergente proliferação de dispositivos vestíveis capazes de monitorar uma gama extensa de dados biométricos representa uma revolução na coleta de dados de saúde.

Esses dispositivos já capturam, em tempo real, informações vitais como frequência cardíaca, pressão arterial, níveis de oxigênio no sangue, padrões de sono e atividade física, entre outros. A integração desses dados em modelos preditivos amplia exponencialmente nossa capacidade de entender os impactos do consumo de álcool e tabaco no bem-estar individual e na saúde pública em geral.

Com o avanço contínuo da tecnologia e a expansão da Internet das Coisas (IoT), os dispositivos vestíveis se tornarão ainda mais sofisticados, aferindo uma quantidade ainda maior de dados biológicos com precisão clínica. Esta evolução promete revolucionar a maneira como monitoramos a saúde em nível individual e populacional, permitindo intervenções mais rápidas, personalizadas e baseadas em dados concretos.

Portanto, a relevância deste estudo é multifacetada: ele não só contribui para a literatura científica atual, fornecendo insights valiosos sobre comportamentos de saúde, mas também estabelece uma base sólida para futuras pesquisas e desenvolvimentos tecnológicos que irão beneficiar a saúde pública em uma escala sem precedentes.

A. Visão Geral do Dataset

O dataset "Smoking and Drinking" inclui informações sobre características físicas e biométricas, além do status de fumante e bebedor dos indivíduos. Este dataset oferece uma oportunidade para explorar como diferentes fatores, como idade, peso, altura e variáveis biométricas, podem estar relacionados com o comportamento de fumo e consumo de álcool.

B. Identificação de Características Relevantes através da PCA

Durante a análise dos dados, aplicamos a técnica de Análise de Componentes Principais (PCA) para reduzir a dimensionalidade do conjunto de dados e identificar as características mais relevantes que contribuem para a variabilidade observada. A PCA transforma o conjunto de dados original, possivelmente correlacionado, em um conjunto de valores de componentes principais linearmente descorrelacionados. As características mais relevantes identificadas pela PCA são descritas a seguir:

1) **Idade (age)**: Conforme esperado, a idade é um fator preditivo significativo para várias incidências de saúde.
2) **Índice de Função Hepática (LIVER_FUNCTION_INDEX)**: Este índice é essencial na avaliação da saúde hepática.

3) **Enzimas Hepáticas (SGOT/ALT, GGT, AST)**: Níveis elevados dessas enzimas podem indicar dano hepático.
4) **Parâmetros Lipídicos (cholesterol total, triglicerídeos, HDL, LDL)**: Têm forte associação com doenças cardiovasculares.
5) **Índice de Risco Cardiovascular (CV_RISK, CARDIOVASCULAR_HEALTH_INDEX)**: Combinam diversos fatores para avaliar o risco de doença cardíaca.
6) **Índice de Massa Corporal (BMI_Cat)**: Classificações de peso corporal que são indicativos de saúde geral.
7) **Status de Fumante (SMK_CAT, smk_stat_type)**: O tabagismo é um fator de risco conhecido para muitas patologias.
8) **Função Renal (KIDNEY_FUNCTION_INDEX, creatinina sérica)**: Indica a saúde renal e a eficácia da filtração.
9) **Hemoglobina e Hemoglobina A1c**: Importantes para diagnóstico de anemia e monitoramento de diabetes.
10) **Avaliação Auditiva e Visual (AVG_SIGHT, AVG_HEAR)**: Alterações podem ser indicativas de condições crônicas ou exposição a riscos.

Através da PCA, foi possível simplificar a complexidade do conjunto de dados, realçando as características mais informativas que devem ser consideradas em futuras análises e desenvolvimento de modelos preditivos.

C. Fonte dos Dados

Os dados são oriundos do Kaggle, uma plataforma popular para competições de ciência de dados e aprendizado de máquina. O dataset está disponível em: https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset.

IV. Metodologia

A. Pré-Processamento dos Dados

O dataset foi inicialmente submetido a um processo de limpeza e pré-processamento. Este processo incluiu as seguintes etapas:

- **Imputação de Valores Faltantes**: Utilizamos a estratégia de imputação média para tratar valores faltantes nas variáveis numéricas.
- **Normalização**: As características numéricas foram normalizadas para ter uma média de zero e desvio padrão de um, usando o `StandardScaler` do scikit-learn.
- **Codificação de Variáveis Categóricas**: As variáveis categóricas, incluindo o sexo do indivíduo e os rótulos dos datasets, foram codificadas numericamente.

B. Divisão do Dataset

O dataset foi dividido em conjuntos de treino e teste usando a função `train_test_split` do scikit-learn, com 80% dos dados destinados ao treinamento e 20% para teste.

C. Modelos de Aprendizado de Máquina

Foram selecionados diversos modelos de aprendizado de máquina para análise e comparação. Os modelos escolhidos foram:

D. K-Nearest Neighbors (KNN)

O K-Nearest Neighbors (KNN) é um algoritmo de aprendizado de máquina supervisionado, predominantemente utilizado para propósitos de classificação. Este algoritmo é amplamente utilizado para a previsão de doenças, destacando-se como uma das formas mais simples e adaptáveis de algoritmos de aprendizado de máquina [20]. O KNN prediz a classificação de dados não rotulados levando em consideração as características e rótulos dos dados de treinamento. Geralmente, ele classifica conjuntos de dados usando um modelo de treinamento semelhante à consulta de teste, considerando os k pontos de dados de treinamento mais próximos (vizinhos) que estão mais próximos da consulta que está sendo testada. Finalmente, o algoritmo realiza uma regra de votação majoritária para finalizar a classificação [21].

1) Funcionamento do KNN: O KNN é simples em seus cálculos e operações. Ele oferece opções para ser modificado em vários aspectos para diminuir suas limitações e desafios e aumentar sua precisão e aplicabilidade em uma variedade mais ampla de conjuntos de dados. O algoritmo clássico do KNN sofre de várias limitações, como ser imparcial a todos os seus vizinhos dependentes da classificação, falta de recursos de cálculo de distância entre pontos de dados e considerar recursos desnecessários do conjunto de dados [22].

E. Random Forest

Random Forest é um algoritmo de aprendizado de máquina supervisionado conhecido por sua robustez e eficácia em tarefas de classificação e regressão. Ele constrói um ”floresta” que é um conjunto de árvores de decisão, geralmente treinadas com o método de ”bagging”. A ideia básica por trás desse algoritmo é combinar várias árvores de decisão para obter um modelo mais poderoso e preciso [12].

1) Funcionamento do Random Forest: No Random Forest, cada árvore na floresta é construída a partir de uma amostra de dados com substituição (conhecido como ”bootstrap sample”). Em cada nó da árvore, um subconjunto aleatório de recursos é selecionado para a divisão. Esse processo de seleção aleatória de recursos e amostras resulta em uma coleção de árvores com variações consideráveis, mas que são capazes de operar como um conjunto forte e unificado [13].

Uma das principais vantagens do Random Forest é sua capacidade de lidar com conjuntos de dados grandes e complexos, mantendo a capacidade de mitigar o overfitting. Isso se deve ao fato de que o aumento do número de árvores na floresta tende a melhorar a robustez do modelo sem o risco de overfitting. Além disso, o Random Forest pode lidar eficazmente com dados ausentes e é capaz de manter uma precisão aceitável, mesmo quando uma grande proporção dos dados está ausente [14].

2) Aplicações do Random Forest: O Random Forest tem sido amplamente utilizado em várias áreas. No setor bancário, por exemplo, é utilizado para identificar clientes leais e detectar fraudes. Na medicina, ajuda na identificação da combinação correta dos componentes para validação de medicamentos e na análise de registros médicos para diagnóstico de doenças.

No mercado de ações, é usado para prever o comportamento das ações e a perda ou lucro esperados. No e-commerce, contribui para o mecanismo de recomendação, ajudando a prever a probabilidade de os clientes gostarem dos produtos recomendados [15].

F. Multi-Layer Perceptron (MLP)

O Multi-Layer Perceptron (MLP) é uma rede neural artificial do tipo feedforward, que consiste em neurônios totalmente conectados com funções de ativação não lineares. O MLP é organizado em pelo menos três camadas: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. É notável por sua capacidade de distinguir dados que não são linearmente separáveis [16].

1) Arquitetura do MLP: O MLP é composto por três ou mais camadas de nós ativadores não lineares. Cada nó em uma camada está conectado a todos os nós na camada seguinte com um certo peso. Essa estrutura totalmente conectada permite que o MLP capture relações complexas e não lineares nos dados [17].

2) Aprendizado no MLP: O aprendizado no MLP ocorre ao ajustar os pesos das conexões após o processamento de cada dado, com base na quantidade de erro na saída em comparação com o resultado esperado. Esse processo é um exemplo de aprendizado supervisionado e é realizado por meio do método de backpropagation, uma generalização do algoritmo de mínimos quadrados em perceptrons lineares [18].

Durante o treinamento, o algoritmo de descida de gradiente é utilizado para ajustar os pesos, de modo a minimizar o erro em toda a saída para cada ponto de dados. O ajuste dos pesos é feito de forma a convergir rapidamente para uma resposta, sem oscilações [19].

Support Vector Machine (SVM): Um modelo poderoso para classificação, especialmente eficaz em espaços de alta dimensão. O SVM é amplamente utilizado devido à sua capacidade de criar hiperplanos ótimos em um espaço de alta dimensão para classificar os dados. Este modelo é particularmente útil em situações onde a clareza da margem de separação entre diferentes classes é de importância crítica.

Cada modelo oferece uma abordagem única para a classificação e análise de dados. A escolha desses modelos para o estudo foi baseada em suas forças e capacidades individuais para lidar com os desafios apresentados pelo conjunto de dados em questão. A compreensão profunda de cada modelo e suas variantes é crucial para aplicar eficazmente técnicas de aprendizado de máquina em problemas do mundo real, como a previsão de comportamentos de saúde.

G. Treinamento e Avaliação dos Modelos

Os modelos foram treinados usando os conjuntos de treino correspondentes e avaliados com base em sua acurácia no conjunto de teste. Além disso, para cada modelo, foram calculadas e analisadas as matrizes de confusão para avaliar o desempenho em diferentes classes.

H. Execução Paralela

Para otimizar o processo de treinamento, os modelos foram treinados em paralelo utilizando o `ThreadPoolExecutor` do Python, permitindo que múltiplos modelos fossem treinados simultaneamente.

V. Implementação

A. Configuração de Treinamento

Os modelos de aprendizado de máquina foram treinados utilizando o conjunto de dados de saúde previamente processado e normalizado. As seguintes estratégias foram empregadas no treinamento:

- **K-Nearest Neighbors (KNN)**: Treinado sem ajuste de hiperparâmetros significativo, utilizando o valor padrão de 5 vizinhos mais próximos.
- **Random Forest**: Utilizado com 100 árvores de decisão. Não foram realizados ajustes de hiperparâmetros adicionais.
- **Multi-Layer Perceptron (MLP)**: Uma rede neural com uma camada oculta de 100 neurônios. O treinamento foi monitorado com a opção de parada antecipada para prevenir overfitting.
- **Support Vector Machine (SVM)**: Devido ao tempo de treinamento extensivo, especialmente evidente em grandes conjuntos de dados, não foi possível concluir o treinamento deste modelo dentro de um prazo razoável.

Não foi aplicada a técnica de validação cruzada, dada a natureza exploratória inicial do projeto e a ênfase na comparação direta do desempenho dos modelos.

B. Desafios Enfrentados

Durante a implementação, enfrentamos desafios significativos, particularmente relacionados ao tempo de treinamento de alguns modelos. O mais notável foi com o modelo SVM, que, devido à sua complexidade computacional, exigiu um tempo de treinamento consideravelmente longo. Após dois dias de treinamento sem conclusão, a decisão foi tomada para abortar este processo. Este desafio ressalta a importância de considerar a eficiência computacional dos modelos em aplicações práticas, especialmente ao lidar com grandes conjuntos de dados. Além disso, a execução de múltiplos modelos em paralelo foi implementada para otimizar o tempo total de processamento.

VI. Resultados e Discussão

A. Avaliação dos Modelos

Os modelos foram avaliados com base em suas acurácias e matrizes de confusão. As acurácias obtidas foram as seguintes:

• MLPClassifier para DRK: 0.74 • RandomForestClassifier para DRK: 0.73 • KNeighborsClassifier para DRK: 0.68 • MLPClassifier para SMK: 0.70 • RandomForestClassifier para SMK: 0.69 • KNeighborsClassifier para SMK: 0.65

VII. Resultados e Discussão

A. Matrizes de Confusão

As matrizes de confusão para cada modelo são apresentadas a seguir.

TABLE I
TABELA 1: KNEIGHBORSCLASSIFIER - DRK

Previsão	Não Bebedor	Bebedor
Real: Não Bebedor	66265	33330
Real: Bebedor	30018	68657

TABLE II
TABELA 2: RANDOMFORESTCLASSIFIER - DRK

Previsão	Não Bebedor	Bebedor
Real: Não Bebedor	72065	27530
Real: Bebedor	26300	72375

TABLE III
TABELA 3: MLPCLASSIFIER - DRK

Previsão	Não Bebedor	Bebedor
Real: Não Bebedor	73720	25875
Real: Bebedor	26589	72086

TABLE IV
TABELA 4: MLPCLASSIFIER - SMK

Previsão	0	1	2
Real: 0	98564	11573	10445
Real: 1	7338	16089	11492
Real: 2	9378	8992	24399

B. Comparação de Modelos

Para avaliar os resultados das matrizes de confusão apresentadas, é importante entender o que cada parte da matriz representa. Em uma matriz de confusão, as colunas representam as classes reais, enquanto as linhas representam as classes previstas pelo modelo. Os valores dentro da matriz indicam o número de previsões para cada combinação de classe real e prevista. Por exemplo, um valor na célula correspondente à classe real "Não Bebedor" e à classe prevista "Não Bebedor" indica o número de verdadeiros negativos, ou seja, instâncias corretamente identificadas como "Não Bebedor".

Vamos analisar as matrizes de confusão fornecidas:

KNeighborsClassifier - DRK: Esta matriz mostra um equilíbrio razoável entre as previsões para bebedores e não bebedores, embora com um número significativo de falsos positivos e negativos.

RandomForestClassifier - DRK: Esta matriz apresenta uma melhoria em relação ao KNeighborsClassifier, com números mais altos de verdadeiros positivos e verdadeiros negativos.

MLPClassifier - DRK: Similar ao RandomForest, esta matriz mostra um bom desempenho com números elevados de previsões corretas tanto para bebedores quanto para não bebedores.

TABLE V
TABELA 5: KNEIGHBORSCLASSIFIER - SMK

Previsão	0	1	2
Real: 0	102512	10057	8013
Real: 1	15033	12099	7787
Real: 2	17269	10845	14655

TABLE VI
TABELA 6: RANDOMFORESTCLASSIFIER - SMK

Previsão	0	1	2
Real: 0	99745	9416	11421
Real: 1	9730	13058	12131
Real: 2	10398	7734	24637

MLPClassifier - SMK: Esta matriz, com três classes, revela uma distribuição mais complexa. O modelo parece ter um desempenho melhor em identificar a classe "0", mas enfrenta dificuldades com as classes "1" e "2".

KNeighborsClassifier - SMK: Aqui, também observamos um desempenho melhor na classe "0", mas com problemas nas classes "1" e "2".

RandomForestClassifier - SMK: Esta matriz mostra um desempenho similar ao KNeighborsClassifier para a classe "0", mas com ligeiras melhorias nas classes "1" e "2".

VIII. Análise dos Resultados

A. Avaliação de Padrões de Comportamento e Implicações para Saúde Pública

As matrizes de confusão geradas a partir dos modelos de classificação fornecem uma análise detalhada dos padrões de consumo de álcool e tabagismo. Esta análise revela não apenas a prevalência desses comportamentos na população estudada, mas também oferece insights sobre a eficácia das políticas de saúde pública atuais. A capacidade de identificar com precisão diferentes grupos - bebedores, não bebedores, e diversos níveis de tabagismo - permite uma avaliação mais refinada das necessidades e dos riscos associados a cada grupo. Estes resultados são fundamentais para orientar o desenvolvimento de estratégias de intervenção e prevenção mais direcionadas e eficazes.

B. Aplicação do Aprendizado de Máquina e Contribuições para Pesquisas Futuras

A utilização de técnicas de aprendizado de máquina na análise de dados de saúde pública representa um avanço significativo na pesquisa comportamental. Os modelos de classificação aplicados demonstram uma capacidade notável de processar e interpretar grandes volumes de dados, oferecendo uma compreensão mais profunda dos padrões comportamentais. Além disso, os resultados obtidos servem como uma base valiosa para pesquisas futuras, proporcionando um ponto de partida para estudos mais detalhados sobre os fatores associados ao consumo de álcool e tabagismo e para o desenvolvimento de intervenções mais específicas e eficientes.

REFERENCES

[1] I. Goodfellow, Y. Bengio, e A. Courville, "Deep Learning," MIT Press, 2016.

[2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] T. Cover e P. Hart, "Nearest neighbor pattern classification," IEEE transactions on information theory, vol. 13, no. 1, pp. 21–27, 1967.

[4] T. Hastie, R. Tibshirani, e J. Friedman, "The Elements of Statistical Learning," Springer Series in Statistics, 2009.

[5] Y. LeCun, Y. Bengio, e G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[6] D. Keim, et al., "Visual Analytics: Definition, Process, and Challenges," Lecture Notes in Computer Science, vol. 4950, 2008.

[7] M. Uddin et al., "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, 2019. Disponível em: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8

[8] M. Uddin et al., "Applications of machine learning in disease pre-screening: A review," Journal of Advanced Research, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S2090123220300540

[9] D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement," International Journal of Surgery, 2010. Disponível em: https://www.sciencedirect.com/science/article/pii/S1743919110007716

[10] A. Rajkomar et al., "Machine learning in medicine," New England Journal of Medicine, 2019. Disponível em: https://www.nejm.org/doi/full/10.1056/NEJMra1814259

[11] K. Jordan, "Machine Learning: An Introduction," 2020. Disponível em: https://www.ibm.com/cloud/learn/machine-learning

[12] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[13] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

[14] C. Strobl et al., "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, no. 1, p. 25, 2007.

[15] B. Kursa et al., "Robustness of Random Forest-based gene selection methods," BMC Bioinformatics, vol. 15, no. 1, p. 8, 2014.

[16] R. Rojas, "Neural Networks: A Systematic Introduction," Springer-Verlag, Berlin, Heidelberg, 1996.

[17] S. Haykin, "Neural Networks and Learning Machines," 3rd ed., Pearson, 2009.

[18] Y. LeCun et al., "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[19] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016.

[20] M. A. Hall et al., "The WEKA data mining software: an update," SIGKDD Explorations, vol. 11, no. 1, 2009.

[21] T. M. Cover e P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[22] A. Hassanat, "Choosing mutation and crossover ratios for genetic algorithms—A review with a new dynamic approach," Information, vol. 6, no. 4, pp. 794–818, 2015.

[23] A. B. Hassanat et al., "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach," International Journal of Computer Science and Information Security, vol. 13, no. 8, 2015.

[24] S. A. Hassanat et al., "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Scientific Reports, vol. 11, 2021.