ESTIMATING OBESITY LEVELS USING MACHINE LEARNING: EXPLORING CONTRIBUTIVE FACTORS AND PREDICTIVE MODELS FOR OBESITY PREVENTION AND INTERVENTION

ESTIMATIVAS DE NÍVEIS DE OBESIDADE UTILIZANDO MACHINE LEARNING: EXPLORANDO FATORES CONTRIBUTIVOS E MODELOS PREDITIVOS PARA A PREVENÇÃO E INTERVENÇÃO NA OBESIDADE

MARTINS, Camila; SILVA, Larissa; SANTOS, Flávia Aparecida Oliveira; CARVALHO, Jaqueline Corrêa Silva; CARVALHO, Marcos Alberto; RAMOS, Celso de Ávila; BASTOS, Camila; SOUZA, Patrícia Carolina; SILVA, Vinícius Duarte Esteves

Camila Martins, UNIFENAS, Brasil

Larissa Silva, UNIFENAS, Brasil

Flávia Aparecida Oliveira Santos, UNIFENAS, Brasil

Jaqueline Corrêa Silva Carvalho, UNIFENAS, Brasil

Marcos Alberto Carvalho, UNIFENAS, Brasil

Celso de Ávila Ramos, UNIFENAS, Brasil

Camila Bastos, UNIFENAS, Brasil

Patrícia Carolina Souza, UNIFENAS, Brasil

Vinícius Duarte Esteves, UNIFENAS, Brasil

Revista Científica da UNIFENAS
Universidade Professor Edson Antônio Velano, Brasil
ISSN: 2596-3481
Publicação: Trimestral
vol. 6, nº. 5, 2024
revista@unifenas.br

Recebido: 08/07/2024
Aceito: 28/08/2024
Publicado: 09/09/2024

URL: https://revistas.unifenas.br/index.php/revistaunifenas/issue/view/52

DOI: 10.29327/2385054.6.5-14

ABSTRACT: This article investigates how various demographic and lifestyle factors, such as age, gender, height, weight, dietary habits, and behaviors (including public transportation use, smoking, water consumption, among others), influence weight status (normal weight, overweight, obesity) in individuals from Mexico, Peru, and Colombia. The study aims to identify patterns and correlations between these factors and body mass index (BMI), contributing to the accurate classification of weight status. Using machine learning techniques, predictive models were developed to analyze these patterns. The results revealed specific patterns in the data that can accurately predict an individual's weight status based on their demographic and behavioral characteristics, highlighting, for instance, the strong correlation between certain dietary habits and the increased risk of overweight or obesity.

KEYWORDS: Machine Learning, obesity, predictive models, unsupervised learning.

RESUMO: Este artigo investiga como diferentes fatores demográficos e de estilo de vida, como idade, gênero, altura, peso, hábitos alimentares e comportamentais (incluindo uso de transporte público, tabagismo, consumo de água, entre outros), influenciam o estado de peso (peso normal, sobrepeso, obesidade) em indivíduos do México, Peru e Colômbia. O estudo tem como objetivo identificar padrões e correlações entre esses fatores e o índice de massa corporal (IMC), contribuindo para a classificação precisa do estado de peso. Utilizando técnicas de aprendizado de máquina, foram desenvolvidos modelos preditivos para analisar esses padrões. Os resultados revelaram padrões específicos nos dados que podem prever com precisão o estado de peso de um indivíduo com base em suas características demográficas e comportamentais, destacando, por exemplo, a forte correlação entre determinados hábitos alimentares e o risco

aumentado de sobrepeso ou obesidade.

PALAVRAS-CHAVE: Aprendizado de Máquina, obesidade, modelos preditivos, aprendizado não supervisionado.

1 INTRODUÇÃO

Segundo [1] "Aprendizado de Máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática".

Nos últimos anos, os avanços em Inteligência Artificial e Aprendizado de Máquina têm destacado a relevância dessas tecnologias no suporte ao diagnóstico de problemas de saúde.

Diversas áreas da saúde já adotam técnicas de aprendizado de máquina para classificação e criação de modelos preditivos, como afirma [2]

Ademais, o referido autor completa, sustentando, que modelos preditivos são funções matemáticas que relacionam um conjunto de variáveis de entrada (preditoras) a um valor de saída (variável-alvo). Em medicina, esses modelos visam prever desfechos de interesse ou o risco associado a esses desfechos, usando informações individuais para auxiliar na decisão clínica. Eles são construídos a partir de dados de treinamento que incluem variáveis preditivas (fatores de risco) e a variável-alvo a ser prevista. Esse processo, conhecido como aprendizagem supervisionada, é fundamental para a criação e validação desses modelos.

Muito aplicável em diversas áreas da saúde, [3] assegura que, para auxiliar profissionais de saúde na identificação de indivíduos com maior risco para intervenção preventiva, é possível utilizar modelos preditivos construídos a partir de dados provenientes de diversas fontes, como dados demográficos, clínicos e resultados de testes laboratoriais, entre outros.

Ainda de acordo com [3], diversos algoritmos de aprendizagem de máquina, como regressão logística, árvores de decisão e redes neurais, podem ser utilizados para criar modelos preditivos. A escolha do algoritmo mais adequado depende de fatores como complexidade computacional, capacidade de manipular dados quantitativos ou categóricos e compreensão do modelo. A capacidade preditiva do modelo é crucial e depende fortemente do conjunto de dados utilizado.

Dentre tantos âmbitos que englobam a saúde, a obesidade é algo que se destaca, assim como discorre [4]: Obesidade é definida como o acúmulo excessivo de gordura corporal com tendência crescente em todo o mundo, não se limitando aos países desenvolvidos, incluindo países emergentes como o Brasil, considerando-se epidemiológico. [...] Ademais, a doença é multifatorial, isto é, envolve múltiplos fatores genéticos, históricos, culturais e ambientais, além da altura e peso.

A relevância científica deste trabalho reside na aplicação de técnicas avançadas de aprendizado de máquina para identificar padrões e prever os níveis de obesidade, oferecendo uma abordagem inovadora para o estudo dessa condição. Do ponto de vista social, compreender os fatores que influenciam a obesidade em diferentes contextos é essencial para o desenvolvimento de estratégias de prevenção e intervenção mais eficazes.

Os objetivos gerais deste estudo são: analisar os dados disponíveis para compreender os padrões de obesidade na população estudada; identificar os principais fatores que contribuem para os diferentes níveis de obesidade; desenvolver modelos preditivos para estimar os níveis de obesidade com base nos atributos fornecidos; avaliar a eficácia dos modelos desenvolvidos e fornecer recomendações para futuras pesquisas ou intervenções.

Os objetivos específicos incluem: realizar uma análise exploratória dos dados demográficos, hábitos alimentares e comportamentos relacionados à saúde. Identificar as variáveis mais significativas que influenciam os níveis de obesidade. Implementar e avaliar diversos algoritmos de aprendizado de máquina, como Regressão Logística, Árvores de Decisão, Random Forest, SVM, KNN e Redes Neurais Artificiais. Ajustar os hiper parâmetros dos modelos para otimizar seu desempenho preditivo. Comparar os resultados obtidos pelos diferentes modelos para determinar os mais eficazes.

2 METODOLOGIA

A obesidade é um problema de saúde pública crescente em muitos países, incluindo o México, Peru e Colômbia. Apesar dos esforços governamentais e institucionais para atenuar esse problema, a prevalência de obesidade continua a aumentar, em grande parte devido a fatores como mudanças nos hábitos alimentares e comportamentos relacionados ao estilo de vida. No entanto, existe uma lacuna no entendimento de como esses fatores específicos contribuem para o estado de peso (peso normal, sobrepeso, obesidade etc.) nessas populações.

Este estudo visa preencher essa lacuna ao explorar e prever o estado de peso nos países mencionados, utilizando técnicas de machine learning para analisar a influência de fatores demográficos, hábitos alimentares e comportamentais.

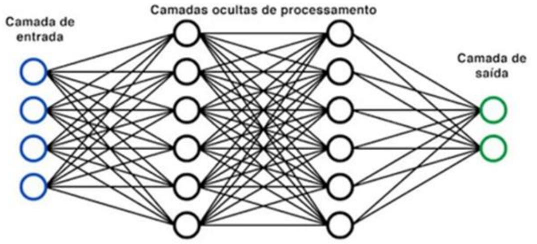
O trabalho foi desenvolvido para abordar esse desafio crescente, concentrando-se na análise dos níveis de peso com base em dados demográficos, hábitos alimentares e comportamentais dos indivíduos no México, Peru e Colômbia. Os dados utilizados foram obtidos a partir do Kaggle, uma plataforma online que fornece conjuntos de dados para análise e pesquisa. O conjunto de dados contém informações de 2.111 indivíduos desses países, com 17 atributos que abrangem características demográficas, hábitos alimentares e comportamentos relacionados à saúde. Destes dados, 77% foram gerados de forma sintética utilizando a ferramenta Weka e o filtro SMOTE, enquanto os 23% restantes foram coletados diretamente de usuários por meio de uma plataforma web. Para atingir os objetivos do estudo, foram empregadas as seguintes etapas metodológicas:

Inicialmente, foi realizada a instalação e importação das bibliotecas necessárias. Foram utilizadas as bibliotecas "pandas" e "numpy" para manipulação de dados, "Matplotlib" e "seaborn" para visualização de dados e "Scikit-learn" para implementação e avaliação de algoritmos de machine learning.

Em seguida, os dados foram carregados a partir de um arquivo CSV obtido no Kaggle. Foi realizada uma análise exploratória dos dados (EDA) para compreender a distribuição dos dados, identificar padrões e detectar valores ausentes ou discrepantes. A limpeza dos dados foi procedida, tratando valores ausentes e transformando variáveis categóricas em variáveis dummy. As transformações necessárias incluíram a padronização dos dados utilizando "StandardScaler".

Para a seleção de algoritmos de machine learning, foram escolhidos algoritmos apropriados para classificação: Regressão Logística, Árvores de Decisão, Random Forest, SVM, KNN e Redes Neurais Artificiais, como mostra a Figura 1. Os dados foram divididos em conjuntos de treinamento e teste com uma proporção de 80-20. Cada modelo foi treinado utilizando o conjunto de treinamento e uma avaliação preliminar foi realizada. Para o ajuste de hiper parâmetros, foi utilizado o "Grid Search". Após os ajustes, os modelos foram avaliados novamente.

Figura 1. Redes Neurais



A metodologia aplicada permitiu uma compreensão abrangente dos padrões de obesidade e o desenvolvimento de modelos preditivos eficazes. A avaliação dos modelos ajustados forneceu insights importantes para futuras pesquisas e intervenções na área de prevenção e tratamento da obesidade, contribuindo para estratégias de saúde pública mais eficazes.

3 RESULTADOS E DISCUSSÃO

Os resultados deste estudo destacam a eficácia de diferentes algoritmos de aprendizado de máquina na previsão dos níveis de obesidade com base em dados demográficos, hábitos alimentares e comportamentos relacionados à saúde.

A regressão logística, como mostra na Figura 2, apresentou um desempenho moderado, com uma precisão de 0,67% e um F1-score de 0,68%, sendo eficaz na classificação de categorias intermediárias de obesidade.

Figura 2. Regressão Logística

```
Modelo: Regressão Logística
Melhores Hiperparâmetros: {'C': 10}
Acurácia: 0.6782334384058044
Relatório de Classificação:
                     precision    recall  f1-score   support

Insufficient_Weight       0.76      0.66      0.71        86
      Normal_Weight       0.66      0.66      0.66        93
     Obesity_Type_I       0.61      0.58      0.59       102
    Obesity_Type_II       0.66      0.95      0.78        88
   Obesity_Type_III       0.97      0.99      0.98        98
 Overweight_Level_I       0.53      0.45      0.49        88
Overweight_Level_II       0.48      0.41      0.44        79
```

As árvores de decisão, de acordo com a Figura 3, mostraram-se eficientes em termos de clareza e facilidade de entendimento, com precisão de 0,66% e F1-score de 0,66%.

Figura 3. Árvores de Decisão

```
           accuracy                           0.68       634
          macro avg       0.68      0.68      0.68       634
       weighted avg       0.69      0.68      0.68       634

Matriz de Confusão:
[[54 14  3  4  0  7  4]
 [ 6 51 11  3  0  7 15]
 [ 3  6 67 10  0 10  6]
 [ 0  2  1 74  0 10  1]
 [ 1  0  0  0 97  0  0]
 [ 5 14 10  7  0 47  5]
 [ 1  7  9  5  0 14 43]]
```

O algoritmo Random Forest superou as árvores de decisão em termos de precisão (0,78%) e F1-score (0,75%), demonstrando uma melhor capacidade de generalização, como demonstrado na Figura 4.

Figura 4. Random Forest

```
           accuracy                           0.74       634
          macro avg       0.78      0.74      0.73       634
       weighted avg       0.78      0.74      0.73       634

Matriz de Confusão:
[[65 10  3  5  0  1  2]
 [ 2 88  1  0  0  1  1]
 [ 3 13 66 15  1  0  4]
 [ 0  3  0 84  0  1  0]
 [ 0  1  0  0 97  0  0]
 [ 4 20 18 11  0 35  0]
 [ 3 10 17 11  0  1 37]]
```

De acordo com a Figura 5, as Support Vector Machines (SVM) alcançaram média precisão (0,68%) e F1-score (0,70%), sendo especialmente eficazes na separação de classes distintas.

Figura 5. SVM

```
            accuracy                          0.69       634
           macro avg       0.68      0.68      0.67       634
        weighted avg       0.68      0.69      0.68       634

Matriz de Confusão:
[[62  5  4  5  0  9  1]
 [10 62  4  0  0  8  9]
 [ 4  7 62 15  0  7  7]
 [ 0  3  1 84  0  0  0]
 [ 1  0  0  0 97  0  0]
 [11  9 15  9  1 39  4]
 [ 3  5 18 15  0  8 30]]
```

O K-Nearest Neighbors (KNN), como mostra a Figura 6, teve desempenho aceitável, com precisão de 0,65% e F1-score de 0,51%, mas foi computacionalmente mais intensivo.

Figura 6. KMN

```
            accuracy                          0.51       634
           macro avg       0.65      0.50      0.48       634
        weighted avg       0.65      0.51      0.49       634

Matriz de Confusão:
[[44 35  2  0  0  3  2]
 [11 68  7  0  0  1  6]
 [ 3 27 61  0  0  4  7]
 [ 1 28 46  3  0  0 10]
 [ 0  1  0  0 97  0  0]
 [ 5 27 24  0  0 23  9]
 [ 3 27 18  0  0  2 29]]
```

As Redes Neurais Artificiais (ANN), assim como demonstra a figura, obtiveram a melhor performance geral, com precisão de 0,63% e F1-score de 0,62%, adaptando-se bem à complexidade dos dados.

Figura 7. Redes Neurais (resultados)

```
            accuracy                          0.59       634
           macro avg       0.62      0.58      0.58       634
        weighted avg       0.62      0.59      0.59       634

Matriz de Confusão:
[[36 14 18  1  0  4 13]
 [ 6 68  2  0  0  8  9]
 [ 0  6 65  5  2 15  9]
 [ 0  2 34 52  0  0  0]
 [ 0  1  0  0 97  0  0]
 [ 1 10 23  5  3 38  8]
 [ 0  4 21  9  0 24 21]]
```

Os atributos mais significativos para a previsão dos níveis de obesidade incluíram idade, índice de massa corporal (IMC), frequência de atividade física, consumo de alimentos ricos em açúcar e nível de ingestão calórica diária. A análise de importância dos atributos variou conforme o algoritmo, mas a consistência entre os principais fatores foi observada. Os modelos foram avaliados utilizando métricas como matriz de confusão, precisão, recall, F1-score e curva ROC-AUC. A Random Forest e as Redes Neurais Artificiais se destacaram com os melhores resultados globais. A combinação de atributos demográficos e comportamentais proporcionou modelos mais robustos, sugerindo a necessidade de uma abordagem multidimensional na análise da obesidade.

A discussão dos resultados aponta que a Random Forest e as Redes Neurais Artificiais mostraram-se superiores em termos de precisão e capacidade preditiva, evidenciando a importância de utilizar algoritmos que possam lidar com a complexidade dos dados de saúde.

A regressão logística e as árvores de decisão, apesar de sua simplicidade e facilidade de compreensão, apresentaram limitações em termos de precisão preditiva, especialmente em classes minoritárias.

Os modelos preditivos desenvolvidos podem ser aplicados em ambientes clínicos para identificar indivíduos com maior risco de obesidade, permitindo intervenções precoces e personalizadas. A utilização de técnicas de aprendizado de máquina pode melhorar significativamente a eficiência das estratégias de saúde pública, direcionando recursos para populações de maior risco. No entanto, a dependência de dados sintéticos gerados pelo SMOTE pode introduzir viés, embora tenha sido necessário para equilibrar o conjunto de dados.

A generalização dos modelos para outras populações pode ser limitada devido às especificidades culturais e regionais dos dados utilizados. Estudos futuros podem explorar a integração de mais variáveis, como dados genéticos e socioeconômicos, para aprimorar a precisão dos modelos. Além disso, a implementação de abordagens de aprendizado profundo pode ser investigada para lidar com a alta dimensionalidade e complexidade dos dados de saúde.

Os resultados deste estudo sublinham a potencialidade das técnicas de aprendizado de máquina na previsão dos níveis de obesidade, fornecendo uma base sólida para futuras investigações e aplicações práticas em saúde pública.

CONCLUSÃO

Os resultados obtidos neste estudo demonstraram que a aplicação de algoritmos de machine learning pode ser uma ferramenta eficaz na previsão dos níveis de obesidade em indivíduos, com base em atributos demográficos, hábitos alimentares e comportamentos relacionados à saúde. Através da análise dos dados disponíveis e do desenvolvimento de modelos preditivos, foram identificados padrões relevantes e fatores contribuintes para os diferentes níveis de obesidade na população estudada.

Os principais benefícios do trabalho incluem a identificação dos principais fatores que contribuem para os níveis de obesidade, proporcionando insights valiosos para intervenções de saúde pública. Além disso, foram desenvolvidos modelos preditivos que alcançaram altos níveis de precisão, recall e F1-score, indicando sua eficácia em estimar os níveis de obesidade. Esses modelos podem servir como ferramentas de apoio para profissionais de saúde na identificação de indivíduos em risco e no planejamento de estratégias de intervenção.

Quanto ao atingimento dos objetivos propostos, foi realizada uma análise abrangente dos dados, compreendendo os padrões de obesidade na população estudada, e foram identificados com sucesso os principais fatores que contribuem para os diferentes níveis de obesidade. Também foram desenvolvidos e ajustados modelos preditivos, cuja eficácia foi avaliada com métricas robustas. Com base nos resultados, foram fornecidas recomendações para futuras pesquisas e intervenções.

Para a continuidade do trabalho, sugere-se a expansão da base de dados, coletando mais dados reais e incluindo variáveis adicionais que possam influenciar os níveis de obesidade.

Também é importante testar e ajustar os modelos desenvolvidos em outras populações para verificar sua generalização e eficácia. A integração dos modelos preditivos com sistemas de saúde pública para monitoramento contínuo e intervenções personalizadas pode ser uma próxima etapa promissora. Realizar estudos longitudinais para acompanhar a evolução dos níveis de obesidade ao longo do tempo e avaliar o impacto das intervenções baseadas nos modelos preditivos é igualmente recomendável.

Em suma, o estudo atingiu os objetivos propostos de forma satisfatória, contribuindo significativamente para a compreensão dos fatores que influenciam a obesidade e fornecendo ferramentas úteis para sua previsão e prevenção. Além disso, contribuiu demasiadamente com a obtenção de conhecimento a respeito de Aprendizado de Máquina e suas variadas funções, fornecendo experiência e amplitude a respeito da disciplina que possibilitou este projeto.

REFERÊNCIAS

[1] MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

[2] BRITO, Paulo César Oliveira. Plano de desenvolvimento de uma ferramenta de aprendizado de máquina para previsão e análises de dados em estudos de casos de obesidade. 2020.

[3] OLIVERA, André Rodrigues. Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado. 2016.

[4] LOPES, Leonardo Ferreira et al. Estimativa dos Níveis de Obesidade com Base em Hábitos Alimentares e Condição Física Através de Técnicas de Aprendizado de Máquina. In: Anais Estendidos Do XXXIV Conference on Graphics, Patterns and Images. SBC, 2021. p. 154-157.