Estimativa dos Níveis de Obesidade com Base em Hábitos Alimentares e Condição Física Através de Técnicas de Aprendizado de Máquina

Leonardo Ferreira Lopes[1], Adonias Caetano de Oliveira[1],
Rhyan Ximenes de Brito[1], Saulo Anderson Freitas de Oliveira[2]
and Luiz Torres Raposo Neto[1]
[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Campus Tianguá, CE, Brasil
[2]Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Campus Tauá, CE, Brasil
E-mails: leonnardo.fer@gmail.com, {adonias.oliveira, rhyan.brito, saulo.oliveira, luiz.raposo}@ifce.edu.br

***Abstract*—Obesity is a chronic disease that affects several countries, causing damage such as respiratory and locomotor difficulties, metabolic changes, cardiovascular problems, and even death, in the extreme case. In this perspective, this initial study aims to evaluate the classifiers' performance, namely, Random Forest and Support Vector Machine, when estimating obesity levels, with data from the set "*Estimation of obesity levels based on eating habits and physical condition Data Set*". Under cross-validation and *Hold-Out*, preliminary results indicate an average accuracy with SVM around 87.84% and RF around 95.18%. Furthermore, we noticed that our approach recognizes overweight and obesity cases better, while such cases, in the latest work, are more critically neglected, misclassifying the most severe degree of obesity. Thus, comparing our results with related works, we concluded that the models studied are suitable to the problem, given the achieved results.**

***Resumo*—Obesidade é uma doença crônica que assola diversos países acarretando prejuízos como dificuldade respiratória e do aparelho locomotor, alterações metabólicas, problemas cardiovasculares e até o óbito, no caso extremo. Nessa perspectiva, o objetivo deste estudo inicial é avaliar o desempenho dos classificadores, a saber, Floresta aleatória e Máquina de Vetor de Suporte, ao estimar os níveis de obesidade, com os dados do conjunto "*Estimation of obesity levels based on eating habits and physical condition Data Set*". Sob a validação cruzada e *Hold-Out*, os resultados preliminares indicam uma média de acurácia com SVM em torno de 87,84% e e RF em torno de 95,18%. Além disso, notamos que nossa abordagem foi capaz de melhor reconhecer os casos de sobrepeso e obesidade. Enquanto que nesses casos, o trabalho mais recente negligencia de forma mais crítica, até mesmo classificando de forma incorreta o grau mais severo de obesidade. Assim, comparando nossos resultados com os trabalhos relacionados, conclui-se que os modelos estudados são adequados ao problema haja vista os resultados atigindos.**

I. Introdução

A obesidade é definida como o acúmulo excessivo de gordura corporal [1] com tendência crescente em todo o mundo, não se limitando aos países desenvolvidos, incluindo países emergentes como o Brasil, considerando-se epidemiológico [2].

No Brasil, o percentual dessa doença entre adultos mais que dobrou, sendo 12,2%, entre 2002 e 2003, para 26,8%, em 2019. Foram registrados nesse mesmo período o crescimento de 43,3% para 61,7% entre adultos com excesso de peso segundo dados divulgados pelo IBGE [3].

Apesar do Índice de Massa Corporal (IMC) ser usado para definir o nível de obesidade, ele não leva em consideração a massa muscular, o que pode sugerir incorretamente uma obesidade. Ademais, a doença é multifatorial, isto é, envolve múltiplos fatores genéticos, históricos, culturais e ambientais, além da altura e peso. Todavia, os aspectos mais associados à obesidade estão relacionados com a desproporção entre a prática de exercícios físicos e a dieta calórica aliadas ao sedentarismo [4].

Para auxiliar os profissionais de saúde na classificação da obesidade com mais precisão, a área de Aprendizado de Máquina (AM) fornece uma vasta área de pesquisa com diversas aplicações possíveis, tais como no desenvolvimento de sistemas inteligentes que classificam e segmentam alimentos apresentados em imagens para auxiliar no monitoramento automático da dieta e ingestão nutricional do usuário [5] e no apoio a elaboração de planos nutricionais para controle do peso [6].

O conjunto de dados utilizado neste trabalho está disponível no repositório da UCI *Machine Learning* desde agosto de 2019, possuindo poucas aplicações, tais como o trabalho de [7]. Os dados são compostos por estimativas de níveis de obesidade com base em diversos fatores além de altura e peso, tais como hábitos alimentares associados com a prática constante de atividades físicas e meio de locomoção utilizado pelos indivíduos. A classificação usou como referência os critérios estabelecidos pela Organização Mundial da Saúde (OMS) aliada à Normativa Mexicana sobre obesidade [8].

Diante disso, o objetivo desta proposta é contribuir com um estudo inicial na área de AM aplicando e analisando o desempenho por meio dos classificadores das Florestas aleatórias (*Random Forest*, RF) e Máquinas de Vetor Suporte (*Support Vector Machine*, SVM) no conjunto de dados *"Estimation of obesity levels based on eating habits and physical condition Data Set"*, desenvolvido por [8].

II. Trabalhos Relacionados

Foram selecionados quatro trabalhos sobre aprendizado de máquina para a classificação da obesidade a fim de servirem como embasamento teórico, como referencial de metodologias adequadas e comparação de resultados.

Em [9] é apresentado um prova de conceito usando vários classificadores (análise discriminante, máquinas de vetores de suporte e redes neurais) para reconhecer a obesidade a partir de 18 variáveis dietéticas e de atividades físicas. A validação cruzada de subamostragem aleatória foi usada para medir a precisão da previsão. Os classificadores superaram as regressões logísticas: a análise discriminante quadrática (QDA) classificou corretamente 59% dos casos versus 55% da regressão logística usando dados originais não balanceados; e o SVM de base radial classificou quase 61% dos casos usando dados balanceados, em comparação com a precisão de predição de 59% da regressão logística. Além disso, o SVM de base radial previu ambas as classes (obeso e não obeso) acima do acaso simultaneamente, enquanto alguns outros métodos alcançaram acurácia de predição acima do acaso para apenas uma classe, geralmente em detrimento da outra.

Os autores de [10] desenvolveram um comitê de classificadores em R com interface em Python para a previsão da obesidade usando o IMC, índice de obesidade, idade, peso, altura e gênero. A abordagem aproveita o modelo linear generalizado, florestas aleatórias e mínimos quadrados parciais. Obtiveram, em média, uma acurácia de 89,68% dos valores previstos de obesidade.

Em [8] é descrito como foi elaborado um conjunto de dados com as estimativas de níveis de obesidade, tendo os dados coletados através de um formulário que foi disponibilizado em uma plataforma web para pessoas do México, Peru e Colômbia com idade entre 14 e 61 anos. Ele possui 2111 registros sendo composto por 23% dos dados oriundos da plataforma e os 77% gerados sinteticamente através do Weka[1] com a Técnica de Super-Amostragem Sintética da Minoria (*Synthetic Minority Over-sampling Technique*, SMOTE).

No trabalho [7] é proposto o desenvolvimento de um sistema para predição e detecção dos níveis de obesidade em jovens, utilizando Árvores de decisão (J48), *Naive Bayes* e Regressão Logística (Logística Simples), além de aplicar a metodologia SEMMA[2] com a versão original do conjunto de dados *"Estimation of obesity levels based on eating habits and physical condition Data Set"*, ou seja, sem o balanceamento descrito em [8]. Com o classificador J48 obtiveram a melhor taxa de precisão (97,4%) com base nas métricas: precisão e revocação. Um *software* foi construído para utilizar e treinar o método selecionado, utilizando o Weka.

III. Materiais e Métodos

Nesta seção estão descritas as etapas de coleta e preparação de dados e como foram realizados os experimentos computacionais. Neste trabalho foi utilizada a ferramenta *Scikit-Learn*.

A. Seleção de conjunto de dados

O conjunto de dados *"Estimation of obesity levels based on eating habits and physical condition Data Set"* foi selecionado porque inclui estimativas dos níveis de obesidade em indivíduos dos países do México, Peru e Colômbia, com base em seus hábitos alimentares e condição física.

Os dados são compostos por estimativas de níveis de obesidade com base em diversos fatores além de altura e peso, tal como hábitos alimentares associados com a prática constante de atividades físicas e meio de locomoção utilizado pelos indivíduos. A classificação usou como referência os critérios estabelecidos pela Organização Mundial da Saúde (OMS) aliada à Normativa Mexicana sobre obesidade.

O conjunto possui 17 atributos, detalhados em *Science direct*[3], e 2111 registros que possibilitam a categorização dos indivíduos em Peso Insuficiente (4), Peso Normal (0), Sobrepeso Nível I (1), Sobrepeso Nível II (2), Obesidade Tipo I (3), Obesidade Tipo II (5) e Obesidade Tipo III (6).

Os dados foram coletados através de uma plataforma *web* com uma pesquisa aos quais os usuários responderam um questionário durante 30 dias [8]. Ao final, os autores realizaram o pré-processamento, aplicando o SMOTE a fim de balancear o conjunto de dados, obtendo um total de 2111 registros compostos por 23% reais advindos da pesquisa enquanto que os outros 77% foram gerados a partir do SMOTE, ver Figura 1. Ao final, o conjunto de dados resultante foi disponibilizado no repositório UCI *Machine Learning*.

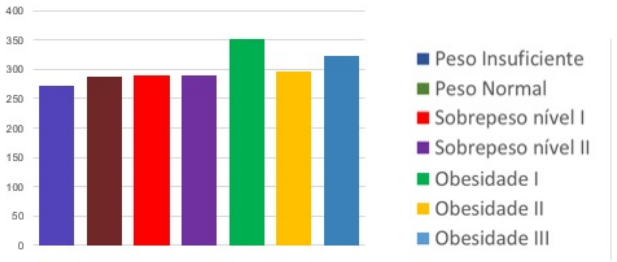

Fig. 1. Base de dados balanceada [8].

B. Pré-processamento de dados

O conjunto de dados possui além de dados numéricos, dados categóricos (quantidade de refeições, ingestão de água, quantidade de vezes que se exercita ao dia, dentre outros) e dados binários em forma de texto. Primeiramente, transformamos estes dados binários em 0 e 1. Já em relação aos dados categóricos, cada categoria foi mapeada em uma nova coluna via técnica *One Hot Encoding*.

Como resultado das etapas citadas, foi obtido um conjunto de dados composto por 27 informações, dos quais, 26 servindo como dados de entrada, e um atributo alvo representando uma da sete possíveis classes.

[1]WEKA é um *software* livre para mineração de dados desenvolvido pela Universidade de Waikato [11].

[2]É um acrônimo que significa Amostrar, Explorar, Modificar, Modelar e Avaliar.

[3]https://www.sciencedirect.com/science/article/pii/S2352340919306985.

C. Experimentos computacionais

Com base nos trabalhos relacionados foi decidido avaliar os modelos RF e SVM (linear). Durante a etapa de experimentos computacionais, foram verificadas várias combinações de parâmetros desses classificadores via validação cruzada *Hold-Out*, 80% (1688) treinamento e 20% (423) teste. Para utilizar o SVM multiclasse, optamos pela versão um contra todos.

IV. Resultados e Discussões

Avaliamos os resultados por meio das métricas Acurácia, Precisão, Revocação, Especificidade e medida-F1, durante as iterações. As médias e desvios padrões de cada métrica são apresentadas na Tabela I.

TABLE I
MÉDIA E DESVIO PADRÃO DE CADA MÉTRICA.

Métricas	**RF**	**SVM**
Acurácia	$94,996\% \pm 0,895\%$	$88,077\% \pm 1,289\%$
Precisão	$95,185\% \pm 0,805\%$	$87,840\% \pm 1,311\%$
Especificidade	$99,166\% \pm 0,149\%$	$98,013\% \pm 0,215\%$
Revocação	$94,834\% \pm 0,912\%$	$87,866\% \pm 1,191\%$
Medida F1	$94,886\% \pm 0,877\%$	$87,596\% \pm 1,275\%$

Conforme pode ser visto na Tabela I, o classificador RF obteve a melhor média em todas as métricas, classificando corretamente em torno de 95% das classes de modo geral, com uma precisão acima de 95% e especificidade acima de 99%. O SVM também apresentou bons resultados, tendo uma acurácia de 88%, além de precisão de 98% e a especifidade 87,9%. De modo geral, pode-se concluir que o modelo RF possui uma ótima medida de exatidão e completude para este problema de classificação, visto que sua medida-F1 foi acima 94%, um ótimo valor de média harmônica ponderada da precisão (exatidão do modelo) e revocação (completude do modelo). O SVM também apresenta boa capacidade de exatidão e completude com base na sua medida-F1 acima de 86,5%.

TABLE II
ANÁLISE COMPARATIVA DAS MÉTRICAS ACURÁCIA E PRECISÃO.

Trabalho	**Técnica**	**Acurácia**	**Precisão**
Kapil Jindal [10]	Comitê – modelo linear generalizado, floresta aleatória e mínimos quadrados parciais.	$89,68\%$	–
De-La-Hoz-Correa [7]	J48	$97,00\%$	$97,40\%$
Selya [9]	SVM RBF	$60,70\%$	–
Nossa proposta	SVM Linear	$88,10\%$	$87,80\%$
Nossa proposta	RF	$95,00\%$	$95,20\%$

A Tabela II apresenta uma comparação entre as métricas acurácia e precisão dos trabalhos relacionados com os desta proposta. Como pode ser visto, os resultados em [9] relatam quase 61% como o melhor resultado dos modelos avaliados. No entanto, nota-se que as demais técnicas empregadas em outros trabalhos, inclusive a nossa, alcançou acurácias maiores (acima de 80%). Adicionalmente, do ponto de vista da saúde, notamos também que os resultados em [10] por não analisarem diversos fatores da obesidade (somente IMC, índice de obesidade, idade, peso, altura e gênero) possuem uma sub-representação do problema. Já a nossa proposta, que emprega todos os atributos, é capaz de atingir bons resultados de desempenho, atingindo entre 80% e 95% nas métricas observadas.

Analisando comparativamente os resultados obtidos neste artigo com os dos trabalhos relacionados (apresentados na Tabela II), conclui-se que, na base dados utilizada, o RF e SVM linear são adequados para estimar os níveis de obesidade em comparação com os utilizados nos trabalhos de [9], [10], em termos de acurácia.

TABLE III
MATRIZ DE CONFUSÃO OBTIDA APLICANDO CLASSIFICADOR RF.

	AP	**NO**	**SBI**	**SBII**	**OBI**	**OBII**	**OBIII**
Abaixo do peso (AP)	52	7	5	0	5	1	0
Normal (NO)	2	46	1	0	0	0	0
Sobrepeso I (SBI)	0	0	53	1	0	0	0
Sobrepeso II (SBII)	0	0	1	63	0	1	0
Obesidade I (OBI)	2	0	0	0	59	0	0
Obesidade II (OBII)	0	0	0	0	0	52	0
Obesidade III (OBIII)	0	0	0	0	0	0	72

As Tabelas III e IV apresentam as matrizes de confusão desta proposta aplicando classificador RF (melhor resultado) e do trabalho de [7], respectivamente. Nota-se que em ambas Tabelas, a quantidade de classes e as quantidades de amostras presentes na matriz de confusão são distintas. Tal situação nos encorajou a fazer uma análise em cima dos percentuais de cada classe para uma análise mais detalhada, principalmente no tocante à identificação das classes que rotulam amostras com algum nível de obsedidade, parte relevante neste trabalho. Assim, derivamos a Tabela V que nos proporciona de forma condensada esses dados de deteção por classe. Adicionalmente, salientamos que as classes de Sobrepeso I (SBI) e sobrepeso II (SBII) foram agrupadas em [7], resultando na classe Sobrepreso (SB).

Como pode ser visto na Tabela V, apesar de uma taxa de acerto alta em função de número absoluto de amostras, como reportado em [7], nota-se que as classes referentes aos nívels mais altos de obesidade são as que possuem menos acurácia na abordagem de De-la-Hoz Manotas [7]. Nós acreditamos que o desbalanceamento prejudica a predição desses níveis. Já em relação à nossa proposta, notou-se que as classes referentes aos nívels mais altos de obesidade foram as detectadas com os maiores percentuais de acurácia.

As piores médias das métricas relatadas neste trabalho foram acima de 86%, enquanto os [9] relatam quase 61% como o melhor resultado dos modelos avaliados. Do ponto de vista da saúde, a base de dados deste trabalho é mais adequada porque analisa diversos fatores da obesidade, uma desvantagem do trabalho de [10], visto que utilizam apenas

TABLE IV
MATRIZ CONFUSÃO DE DE-LA-HOZ MANOTAS [7].

	AP	**NO**	**SB**	**OBI**	**OBII**	**OBIII**
Abaixo do peso (AP)	40	0	0	0	0	0
Normal (NO)	4	360	0	0	0	0
Sobrepeso (SB)	0	0	180	0	0	0
Obesidade I (OBI)	0	0	4	92	0	0
Obesidade II (OBII)	0	0	0	8	20	4
Obesidade III (OBIII)	0	0	0	0	0	0

TABLE V
PERCENTUAIS DE ACURÁRIA POR CLASSE.

Classe	De-la-Hoz Manotas [7]	Nossa proposta
Abaixo do peso	100,00	74,00
Normal	99,00	93,87
Sobrepeso	100,00	97,50[4]
Obesidade I	96,00	97,00
Obesidade II	62,00	100,00
Obesidade III	0,00	100,00

seis fatores. Em relação aos resultados, este trabalho dispõe de modelos com excelentes médias de acurácia.

V. CONCLUSÕES E TRABALHOS FUTUROS

Do ponto de vista de sistema apoio, este trabalho apresenta um estudo preliminar de análise de fatores que visam classificar os níveis de obesidade. Foram avaliados as técnicas das Florestas Aleatórias e Máquinas de Vetor Suporte sob a validação cruzada *Hold-out*. O classificador RF alcançou excelentes médias de acurácia, precisão e revocação próximas de 95%. Já o SVM, obteve médias de acurácia, precisão e revocação próximas de 88%.

Comparando nossos resultados com os trabalhos relacionados, conclui-se que os modelos estudados são adequados ao problema de estimar o nível da obesidade haja vista os resultados atigindos. Em relação a dois trabalhos da literatura, atingimos resultados superiores em relação à acuracia. Adicionalmente, apesar de não termos atingido acurácia superior em relação ao terceiro trabalho, o mais recente, notamos que nossa abordagem foi capaz de melhor reconhecer os casos de sobrepeso e obesidade. Enquanto que nesses casos, o trabalho mais recente negligencia de forma mais crítica, até mesmo classificando de forma incorreta o grau mais severo de obesidade.

Como trabalhos futuros sugere-se analisar mais fatores associados, empregar algoritimos de aprendizagem profunda e disponibilizar um *software* mais completo com interface gráfica. Tudo isso visando auxiliar o diagnóstico médico.

REFERENCES

[1] World Health Organizaion, "Obesity: preventing and managing the global epidemic," 2000.

[2] A. R. d. O. Pinheiro, S. F. T. d. Freitas, and A. C. T. Corso, "Uma abordagem epidemiológica da obesidade," *Revista de Nutrição*, vol. 17, no. 4, pp. 523–533, 2004.

[3] E. C. C. d. S. Bulsing *et al.*, "Asma, índice de massa corporal e sintomas respiratórios: um estudo de base populacional," 2019.

[4] E. N. Wanderley and V. A. Ferreira, "Obesidade: uma perspectiva plural," *Ciência & Saúde Coletiva*, vol. 15, pp. 185–194, 2010.

[5] C. N. Freitas, F. R. Cordeiro, and V. Macario, "Myfood: A food segmentation and classification system to aid nutritional monitoring," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2020, pp. 234–239.

[6] A. M. da Rocha Fernandes, M. B. Pinheiro, and A. G. Nazário, "Aplicação de Algoritmos de Aprendizado de Máquina no Apoio a Elaboração de Planos Nutricionais," *Brazilian Journal of Development*, vol. 6, no. 8, pp. 60 935–60 944, 2020.

[7] A. De la Hoz Manotas, E. De la Hoz Correa, F. Mendoza, R. Morales, and B. Sanchez, "Obesity Level Estimation Software based on Decision Trees," *Journal of Computer Science*, vol. 15, p. 10, 01 2019.

[8] F. M. Palechor and A. de la Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data in Brief*, vol. 25, p. 104344, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340919306985

[9] A. S. Selya and D. Anshutz, "Machine learning for the classification of obesity from dietary and physical activity patterns," in *Advanced Data Analytics in Health*. Springer, 2018, pp. 77–97.

[10] K. Jindal, N. Baliyan, and P. S. Rana, "Obesity prediction using ensemble machine learning approaches," in *Recent Findings in Intelligent Computing Techniques*. Springer, 2018, pp. 355–362.

[11] M. A. H. e. C. J. P. Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 2016. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf