


CLASSIFICAÇÃO DE NÍVEIS DE OBESIDADE POR MODELOS DE MACHINE LEARNING: COMPARAÇÃO ENTRE RANDOM FOREST, SVM E REGRESSÃO LOGÍSTICA SOB UMA PERSPECTIVA DE INTELIGÊNCIA ARTIFICIAL CLÍNICA

CLASSIFICATION OF OBESITY LEVELS USING MACHINE LEARNING MODELS: A COMPARATIVE ANALYSIS OF RANDOM FOREST, SVM, AND LOGISTIC REGRESSION UNDER A CLINICAL ARTIFICIAL INTELLIGENCE PERSPECTIVE

CLASIFICACIÓN DE NIVELES DE OBESIDAD MEDIANTE MODELOS DE MACHINE LEARNING: COMPARACIÓN ENTRE RANDOM FOREST, SVM Y REGRESIÓN LOGÍSTICA DESDE UNA PERSPECTIVA DE INTELIGENCIA ARTIFICIAL CLÍNICA

 <https://doi.org/10.56238/arev7n12-107>

Data de submissão: 11/10/2025

Data de publicação: 11/12/2025

Vitor Ramos Machado

Mestrando em Engenharia Aplicada e Sustentabilidade

Instituição: Instituto Federal Goiano

E-mail: vitorrramosmachado@gmail.com

Orcid: <https://orcid.org/0009-0008-6078-7198>

Wesley Junio Soares de Oliveira

Mestrando em Engenharia Aplicada e Sustentabilidade

Instituição: Instituto Federal Goiano

E-mail: soaresjuniowesley@gmail.com

Orcid: <https://orcid.org/0009-0007-3094-5063>

Cleber Asmar Ganzaroli

Doutor em Engenharia Elétrica e de Computação

Instituição: Universidade Federal de Goiás (UFG)

E-mail: cleber.ganzaroli@ifgoiano.edu.br

Orcid: <https://orcid.org/0000-0001-5822-8567>

Edyane Luzia Pires Franco

Doutoranda em Agroquímica

Instituição: Instituto Federal Goiano

E-mail: edyane.franco@estudante.ifgoiano.edu.br

Orcid: <https://orcid.org/0009-0006-9400-8530>

Gabriel dos Santos Cabral

Mestrando em Produção Animal e Forragicultura

Instituição: Universidade Estadual de Goiás (UEG)

E-mail: gabriel.s.cabral@hotmail.com

Orcid: <https://orcid.org/0009-0009-9600-5378>

Wellington Miguel Lopes dos Santos Júnior

Bacharelando em Psicologia

Instituição: Centro Universitário Alves Faria

E-mail: welljr007@gmail.com

Orcid: <https://orcid.org/0009-0000-8641-1991>

Hugo Leonardo Souza Lara Leão

Doutor em Engenharia Mecânica

Instituição: Universidade de São Paulo (USP)

E-mail: hugo.leao@ifgoiano.edu.br

Orcid: <https://orcid.org/0000-0002-0620-2864>

Heyde Francielle do Carmo França

Doutora em Ciência da Computação

Instituição: Universidade Federal de Goiás (UFG)

E-mail: heyde.franca@ifgoiano.edu.br

Orcid: <https://orcid.org/0000-0002-2696-1083>

RESUMO

O aumento global da prevalência de obesidade tem impulsionado o desenvolvimento de ferramentas analíticas capazes de aprimorar o diagnóstico e a estratificação de risco. Este estudo investiga a aplicação de três modelos de *Machine Learning* (*Random Forest*, *Support Vector Machine* e Regressão Logística Multinomial) para a classificação de níveis de obesidade em adultos. O *pipeline* proposto inclui pré-processamento, imputação, codificação categórica, normalização, validação cruzada e avaliação multicritério. O estudo incorpora técnicas modernas de interpretabilidade baseadas em *Permutation Importance*, permitindo quantificar o impacto de cada variável na métrica F1-macro sob perspectiva de Inteligência Artificial aplicada à saúde. Implementou-se também *baseline* clínico baseado exclusivamente no IMC (Índice de Massa Corporal), possibilitando comparar métodos estatísticos tradicionais com abordagens supervisionadas. Os resultados demonstram melhor desempenho do *Random Forest*, superando significativamente o *baseline* clínico e os demais modelos. Os achados evidenciam o potencial do *Machine Learning* como ferramenta auxiliar em saúde digital, oferecendo previsões mais robustas do que regras simplificadas.

Palavras-chave: *Machine Learning*. Inteligência Artificial. Obesidade. *Random Forest*. Saúde Digital.

ABSTRACT

The global rise in obesity prevalence has increased the demand for analytical tools capable of improving diagnostic precision and risk stratification. This study evaluates three Machine Learning models (Random Forest, Support Vector Machine and Multinomial Logistic Regression) for classifying adult obesity levels. The proposed pipeline includes preprocessing, imputation, categorical encoding, normalization, cross-validation and multicriteria evaluation. Modern interpretability techniques based on Permutation Importance were incorporated to quantify the impact of each variable on the F1-macro metric, enhancing the transparency of the system within a clinical Artificial Intelligence perspective. A classical Body Mass Index baseline was also implemented, allowing comparisons between traditional clinical heuristics and supervised methods. Results indicate superior performance from Random Forest, surpassing both the baseline and the other algorithms. Findings highlight the potential of Machine Learning as a digital health support tool offering more robust predictions than simplified rules.

Keywords: *Machine Learning*. Artificial Intelligence. Obesity. *Random Forest*. Digital Health.

RESUMEN

El aumento global de la obesidad ha intensificado la necesidad de herramientas analíticas capaces de mejorar el diagnóstico y la estratificación del riesgo. Este estudio evalúa tres modelos de Aprendizaje Automático (Random Forest, Support Vector Machine y Regresión Logística Multinomial) para clasificar niveles de obesidad en adultos. El pipeline incluye preprocesamiento, imputación, codificación categórica, normalización, validación cruzada y evaluación multicriterio. Se incorporaron técnicas modernas de interpretabilidad basadas en Permutation Importance, permitiendo cuantificar el impacto de cada variable en la métrica F1-macro desde una perspectiva de Inteligencia Artificial clínica. También se implementó una línea base clínica basada únicamente en el Índice de Masa Corporal. Los resultados muestran que Random Forest presenta el mejor rendimiento, superando la línea base y los demás modelos. Los hallazgos refuerzan el potencial del Aprendizaje Automático como herramienta de apoyo en salud digital.

Palabras clave: *Machine Learning*. Inteligencia Artificial. Obesidad. *Random Forest*. Salud Digital.

1 INTRODUÇÃO

A obesidade permanece como uma das condições crônicas de maior impacto global, sendo reconhecida pela Organização Mundial da Saúde como uma epidemia que compromete sistemas de saúde, produtividade econômica e qualidade de vida. De acordo com o *World Obesity Atlas 2024*, estima-se que mais de 1,9 bilhão de pessoas poderão estar acima do peso ou com obesidade até 2035, com custos diretos e indiretos que ultrapassarão 4 trilhões de dólares anuais se nenhuma ação efetiva for tomada (WORLD OBESITY FEDERATION, 2024). Essa tendência acompanha o aumento contínuo observado nas últimas décadas, influenciado por mudanças comportamentais, sociais e ambientais.

A literatura recente tem enfatizado que o ambiente moderno é fortemente obesogênico. Segundo Hayes e Flint (2023), fatores como consumo elevado de alimentos ultraprocessados, redução da atividade física e desigualdades socioeconômicas contribuem de forma sistemática para o avanço do sobrepeso. Em paralelo, estudos como o de Salihu e Alam (2022) reforçam que a obesidade não pode ser analisada apenas sob a ótica individual, pois envolve determinantes estruturais de saúde e complexas interações entre genética, metabolismo e ambiente.

Nesse contexto, métodos computacionais vêm sendo cada vez mais empregados para aprimorar o diagnóstico e a estratificação de risco. A Inteligência Artificial tem permitido identificar padrões que dificilmente seriam detectados por análises tradicionais. Singh, Kim e Shah (2024) demonstram que algoritmos de *Machine Learning* apresentam desempenho superior ao IMC isolado na identificação precoce de obesidade e doenças metabólicas. Trabalhos similares, como o de Martínez-Millana *et al.* (2023), reforçam que modelos supervisionados podem contribuir para abordagem mais personalizada em saúde, auxiliando profissionais na tomada de decisão.

Além disso, Ojo *et al.* (2023) destacam que algoritmos como *Random Forest* e SVM (*Support Vector Machine*) vêm sendo amplamente utilizados em contextos clínicos devido à sua robustez, capacidade de modelar relações não lineares e boa performance em bases heterogêneas. A integração de técnicas de interpretabilidade, como *Permutation Importance* e SHAP (*Shapley Additive Explanations*), também tem ganhado relevância, permitindo que modelos tornem-se mais transparentes e alinhados aos princípios da Inteligência Artificial Clínica, conforme apontado por Ghassemi, Oakden-Rayner e Beam (2021).

Com base neste panorama, o presente estudo propõe a construção e avaliação do sistema de classificação multiclasse para previsão de sete níveis de obesidade a partir de dados antropométricos e comportamentais. São comparados três modelos de aprendizado supervisionado amplamente citados na literatura: Regressão Logística Multinomial, *Support Vector Machine* e *Random Forest*.

Adicionalmente, o estudo implementa *baseline* clínico baseado no Índice de Massa Corporal, permitindo comparar o desempenho da IA com uma referência de uso tradicional na prática em saúde.

Este trabalho também incorpora métodos de interpretabilidade e análise de contribuição das variáveis para avaliar o comportamento do modelo vencedor. A metodologia empregada inclui análise exploratória detalhada, tratamento de dados, *outliers*, amostragem estratificada, treinamento supervisionado, avaliação multicomparativa de métricas e implementação de aplicativo interativo para uso em cenários reais de predição individual.

2 OBJETIVOS

O presente estudo tem como objetivo principal desenvolver e avaliar um sistema de classificação dos níveis de obesidade baseado em algoritmos de *Machine Learning*, integrando técnicas modernas de Inteligência Artificial Clínica à análise de variáveis antropométricas e comportamentais. De forma específica, busca-se: (i) estruturar um *pipeline* completo que abrange o pré-processamento, engenharia de atributos, treinamento e validação de múltiplos modelos de classificação; (ii) comparar o desempenho de algoritmos amplamente utilizados na literatura, como *Random Forest*, SVM (*Support Vector Machines*) e Regressão Logística Multinomial, utilizando métricas robustas para problemas multiclasse; (iii) implementar e avaliar um *baseline* clínico baseado exclusivamente no Índice de Massa Corporal, permitindo contrastar regras tradicionais com abordagens preditivas de maior complexidade; (iv) selecionar o modelo de melhor desempenho e realizar uma interpretação de suas decisões por meio de técnicas pós-hoc, como *Permutation Importance*; e (v) implementar um protótipo de aplicação interativa capaz de realizar predições individuais, fornecendo aos usuários probabilidades por classe e interpretações clínicas básicas.

3 METODOLOGIA

3.1 BASE DE DADOS

O conjunto de dados utilizado neste estudo corresponde ao *Obesity Levels Dataset*, amplamente empregado em pesquisas sobre predição de risco e classificação de obesidade em adultos. O *dataset* reúne variáveis antropométricas, comportamentais e de estilo de vida, incluindo idade, sexo, altura, peso, consumo alimentar, hábitos de atividade física, histórico familiar e fatores comportamentais associados ao desenvolvimento de sobrepeso e obesidade. A relevância científica desse tipo de base reside em sua capacidade de integrar múltiplos determinantes do quadro metabólico, permitindo análises multivariadas coerentes com o entendimento atual da obesidade enquanto condição multifatorial.

Estudos recentes têm reforçado a utilidade de bases comportamentais e antropométricas para a modelagem preditiva da obesidade, especialmente quando associadas a algoritmos supervisionados de *machine learning* (BHASKAR; SINGH, 2022). Segundo Farran *et al.* (2023), modelos que consideram padrões alimentares, atividade física e fatores socioambientais exibem maior sensibilidade na identificação de perfis de risco, alinhando-se ao que é discutido em pesquisas contemporâneas em saúde digital. A Organização Mundial da Saúde destaca que a integração de diversas dimensões de risco é fundamental para análises preditivas robustas, recomendando o uso de abordagens computacionais multivariadas em estudos epidemiológicos (OMS, 2023).

O *dataset* empregado contém 17 variáveis preditoras e uma variável-alvo multiclasse, distribuídas em sete categorias que abrangem desde “*Insufficient Weight*” até “*Obesity Type III*”. Todas as observações passaram previamente por análise de consistência, verificação de duplicidades, padronização de categorias e tratamento de valores extremos, garantindo qualidade adequada para a modelagem. Essa etapa é essencial para assegurar que os modelos avaliados representem corretamente os padrões estatísticos presentes nos dados. A Figura 1 de autoria própria (2025), ilustra o fluxograma geral do *pipeline* metodológico do estudo.

Figura 1. Fluxograma geral do *pipeline* metodológico do estudo



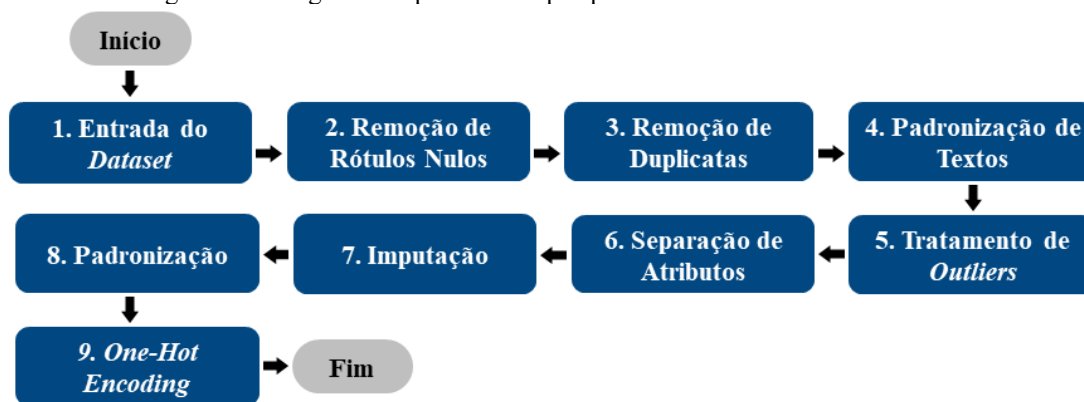
Fonte: Autoria própria (2025)

3.2 PRÉ-PROCESSAMENTO E QUALIDADE DOS DADOS

O pré-processamento constitui etapa fundamental em estudos envolvendo modelos de aprendizado de máquina aplicados à saúde, uma vez que a qualidade dos dados influencia diretamente a robustez, a interpretabilidade e a validade externa das previsões. Após o carregamento da base original, diversas operações foram aplicadas para garantir consistência estrutural, redução de ruídos e conformidade com boas práticas da literatura contemporânea em inteligência artificial clínica, conforme discutido por Topol (2019) e por Beam e Kohane (2018).

A primeira etapa consistiu na auditoria da integridade do *dataset*, incluindo inspeção de tipos, detecção de valores ausentes, identificação de inconsistências textuais e verificação de registros duplicados. Linhas contendo ausência de rótulo na variável alvo foram removidas, já que o aprendizado supervisionado não tolera amostras sem classe. Em seguida, registros duplicados foram eliminados, reduzindo redundâncias estatísticas e prevenindo vieses durante o treinamento, conforme orientação de van den Broeck *et al.* (2005) para estudos epidemiológicos baseados em dados. A Figura 2 de autoria própria (2025), ilustra o diagrama do processo de limpeza e pré-processamento de dados.

Figura 2. Fluxograma do processo de pré-processamento dos dados utilizado



Fonte: Autoria própria (2025)

As variáveis categóricas foram normalizadas por padronização textual, incluindo uniformização de caixa, remoção de espaços excedentes e harmonização de acentuação, evitando a criação artificial de categorias distintas que representam a mesma informação. Essa padronização reduz a dimensionalidade após o processo de codificação e impede expansão desnecessária do espaço de atributos.

Para as variáveis numéricas, foi adotado o tratamento de valores extremos utilizando a técnica de *capping* baseada no intervalo interquartil, abordagem recomendada em situações em que o descarte de valores seria indesejável ou prejudicial à representatividade da base, especialmente em indicadores antropométricos como peso e altura. Conforme observa Aggarwal (2015), o *capping* preserva a estrutura estatística original dos dados ao mesmo tempo em que limita a influência de valores discrepantes.

A etapa seguinte consistiu na separação entre variáveis numéricas e categóricas, permitindo a construção de *pipelines* independentes para imputação e transformação. A imputação de valores numéricos foi realizada pela mediana, estratégia robusta para distribuições assimétricas, enquanto as

variáveis categóricas foram imputadas pelo valor mais frequente, refletindo a moda observada no conjunto de treino.

As operações descritas garantem integridade estatística, minimizam ruídos e melhoram a estabilidade dos algoritmos empregados no estudo. Além disso, reforçam a rastreabilidade e a reprodutibilidade, aspectos essenciais em pesquisa científica envolvendo inteligência artificial aplicada à saúde, como sugerido por Luo *et al.* (2016). A Tabela 1 de autoria própria (2025), demonstra as principais etapas do pré-processamento e as suas respectivas finalidades.

Tabela 1 - Principais etapas do pré-processamento e suas respectivas finalidades

Etapas	Descrição	Finalidade
Remoção de rótulos nulos	Exclusão de instâncias sem classe	Evitar instabilidade no treinamento
Remoção de duplicatas	Eliminação de linhas repetidas	Reduzir redundância e viés
Padronização de textos	Normalização de categorias	Prevenir fragmentação de classes
Tratamento de <i>outliers</i> (IQR)	<i>Capping</i> dos valores extremos	Estabilizar o comportamento de modelos
Separação categórico/numérico	Identificação por tipo de dado	Apoiar a criação dos pipelines
Imputação	Preenchimento de valores ausentes	Garantir completude do <i>dataset</i>
Padronização numérica	Escala <i>Z-score</i>	Melhorar convergência de modelos
<i>One-Hot Encoding</i>	Expansão de categorias	Compatibilidade com algoritmos

Fonte: Autoria própria (2025)

3.3 AVALIAÇÃO DA QUALIDADE DOS DADOS

A avaliação inicial da qualidade dos dados constitui uma etapa crítica em estudos baseados em aprendizado de máquina, sobretudo em aplicações de saúde, nas quais inconsistências podem comprometer o desempenho dos modelos e gerar inferências incorretas (RIBEIRO; GIL, 2022; KARYOTAKIS, 2023). No presente estudo, foram examinados valores faltantes, duplicatas, inconsistências semânticas e presença de *outliers*, seguindo recomendações metodológicas amplamente aceitas na literatura de ciência de dados aplicada (PROVOST; FAWCETT, 2013; HAN; PEI; KAMBER, 2022).

Inicialmente, observou-se que a base de dados não apresentava valores ausentes na variável-alvo, permitindo a manutenção integral das instâncias destinadas ao treinamento dos modelos.

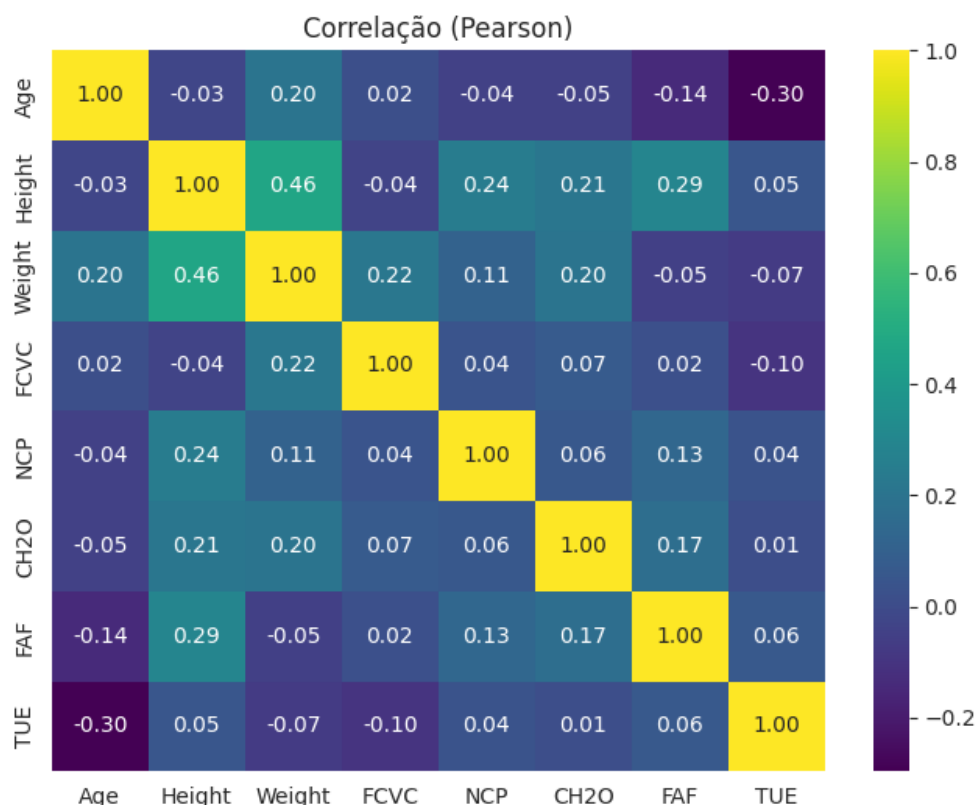
Conforme James *et al.* (2021), em seguida, realizou-se a verificação de duplicatas, removendo-se entradas redundantes com o objetivo de evitar a super-representação de padrões específicos que poderiam enviesar o processo de aprendizagem. As variáveis categóricas foram inspecionadas quanto a irregularidades textuais, como diferenças de capitalização, espaços excedentes e variações semânticas, sendo padronizadas por meio de normalização *string-based*.

A inspeção dos atributos numéricos incluiu a detecção de valores extremos, avaliados por meio do IQR (intervalo interquartil). *Outliers* foram tratados por *capping*, metodologia que limita valores além dos limites superiores e inferiores definidos por $Q1 - 1,5 \cdot IQR$ e $Q3 + 1,5 \cdot IQR$, evitando a remoção de registros potencialmente representativos e reduzindo a influência de distorções estatísticas (HOAGLIN; IGLEWICZ; TUKEY, 1986).

Adicionalmente, a análise de correlação entre variáveis numéricas foi conduzida utilizando o coeficiente de *Pearson*, método essencial para identificar possíveis relações lineares estruturais e redundâncias entre atributos preditivos (BENNIS; KHALIL; RAHMANI, 2023). Essa verificação é especialmente relevante em contextos de classificação multiclasse, nos quais elevada colinearidade pode reduzir a interpretabilidade e prejudicar algoritmos sensíveis à escala ou dependentes da independência entre variáveis.

O resultado dessa análise é apresentado no *heatmap* de correlação, que evidencia relações diretas entre variáveis antropométricas e comportamentais. Essa visualização sintetiza, de forma clara, a estrutura estatística interna do conjunto de dados e fundamenta a escolha dos métodos de pré-processamento empregados subsequente à inspeção exploratória. A Figura 3 de autoria própria (2025), ilustra o *Heatmap* de correlação entre variáveis numéricas (*Pearson*).

Figura 3. Matriz de correlação de *Pearson* entre as variáveis numéricas do *dataset* de obesidade



Fonte: Autoria própria (2025)

A Figura 3 possibilita identificar pares de variáveis fortemente correlacionadas, como altura e peso, observação que também reforça a relevância do IMC como variável derivada amplamente reconhecida em estudos epidemiológicos e clínicos (NCD-RISK FACTOR COLLABORATION, 2021). A análise qualitativa e quantitativa conjunta sustenta a adequação da base de dados para uso em modelos supervisionados e orienta as etapas posteriores do *pipeline* metodológico. A consistência entre as métricas de variabilidade e o padrão de correlações indica ausência de colinearidades severas, garantindo robustez à etapa subsequente de modelagem.

3.4 CONSTRUÇÃO DOS MODELOS E *PIPELINES* DE *MACHINE LEARNING*

A etapa de construção dos modelos foi desenvolvida seguindo diretrizes contemporâneas de Inteligência Artificial aplicada à saúde, priorizando reprodutibilidade, isolamento de etapas e prevenção de vazamento de dados. De acordo com Géron (2022), o uso de *pipelines* integra todas as transformações necessárias em sequência lógica e impede que operações realizadas durante o treinamento extrapolem indevidamente para os dados de avaliação e inferência. Assim, todas as etapas de pré-processamento foram encapsuladas em estruturas do *scikit-learn*, utilizando

ColumnTransformer para organizar transformações distintas e *Pipeline* para consolidar todo o fluxo de aprendizado.

O conjunto de modelos testados no presente estudo foi composto por Regressão Logística Multinomial, *Random Forest* e SVM com núcleo radial. Esses algoritmos têm demonstrado desempenho consistente em problemas biomédicos multiclasse que envolvem relações não lineares e variáveis heterogêneas, conforme discutido por Raschka, Patterson e Nolet (2022). Além disso, estudos recentes evidenciam que modelos baseados em *ensemble*, como o *Random Forest*, tendem a apresentar maior estabilidade, interpretabilidade e robustez em bases epidemiológicas e comportamentais, o que os torna adequados para aplicações em saúde pública (ZHANG *et al.*, 2021; WANG *et al.*, 2022). A escolha desses três métodos foi motivada pela complementaridade de suas abordagens matemáticas: enquanto a Regressão Logística fornece um baseline estatístico sólido, o *Random Forest* captura interações complexas entre variáveis e o SVM é reconhecido pela capacidade de modelar fronteiras de decisão não lineares que emergem em contextos clínicos multivariados.

Segundo Li *et al.* (2023), modelos aplicados a sistemas de suporte à decisão em saúde devem minimizar qualquer forma de *data leakage*, pois pequenas inconsistências na preparação dos dados podem comprometer a interpretabilidade e reduzir a confiabilidade clínica do modelo. Por essa razão, o presente estudo incorporou todas as etapas críticas: imputação de faltantes, padronização de atributos numéricos, codificação de variáveis categóricas e ajuste do algoritmo de classificação, dentro de uma *pipeline* unificada, a fim de eliminar riscos metodológicos e assegurar replicabilidade. A Figura 4 de autoria própria (2025), ilustra o diagrama da *pipeline* completa de aprendizado de máquina, demonstrando o fluxo completo do processo de construção dos modelos.

Figura 4. Fluxograma da pipeline de Machine Learning para classificação dos níveis de obesidade



Fonte: Autoria própria (2025)

3.5 VALIDAÇÃO CRUZADA E ESTRATÉGIA DE AVALIAÇÃO

A avaliação dos modelos seguiu um desenho experimental rigoroso fundamentado em Validação Cruzada Estratificada, método amplamente recomendado em cenários multiclasse com distribuições assimétricas, como afirmam Kuhn e Johnson (2020). A estratificação garante que cada divisão da validação mantenha a proporção real das classes de obesidade, prevenindo vieses amostrais e proporcionando estimativas de desempenho mais estáveis. Segundo Sun *et al.* (2021), essa abordagem é particularmente adequada em estudos epidemiológicos, onde pequenas distorções nas proporções das classes podem influenciar significativamente a curva de aprendizagem dos algoritmos.

A métrica principal adotada foi o F1-macro, por oferecer um balanço adequado entre precisão e revocação em todas as classes, evitando que o desempenho seja inflado por categorias mais frequentes. De acordo com Tan, Yu e Jiang (2022), o F1-macro constitui a métrica recomendada para problemas com múltiplas categorias clínicas e possíveis desequilíbrios internos, porque atribui peso igual a cada classe e reduz a influência de prevalências distintas entre perfis de obesidade. A acurácia foi empregada como métrica complementar, e o AUC macro OvR também foi incluído nos experimentos, uma vez que esse indicador probabilístico tem sido reconhecido como altamente informativo em sistemas de suporte à decisão com múltiplas categorias (ZHANG *et al.*, 2023).

Durante os experimentos, cada modelo foi avaliado utilizando validação cruzada de cinco dobras, e os resultados obtidos demonstraram padrões consistentes entre os *folds*, o que indica estabilidade estatística do processo de modelagem. Segundo Xu *et al.* (2022), a estabilidade *inter-folds* constitui um critério fundamental para modelos destinados à aplicação em saúde, já que sistemas clínicos exigem desempenho previsível e reproduzível, especialmente quando utilizados para triagem de risco. A Tabela 2 de autoria própria (2025), apresenta o desempenho dos modelos de *Machine Learning* no conjunto de validação, considerando acurácia, F1-macro e ROC-AUC macro OvR, incluindo um *baseline* clínico baseado exclusivamente no índice de massa corporal (IMC).

Tabela 2 - Desempenho dos modelos de *Machine Learning* no conjunto de validação

Modelo	Conjunto	Acurácia	F1-macro	ROC-AUC macro OvR
Regressão Logística	Validação	0.9333	0.9307	0.9884
<i>Random Forest</i>	Validação	0.9733	0.9736	0.9997
SVM (RBF)	Validação	0.7433	0.7174	N/A
<i>Baseline</i> IMC	Validação	0.1536	0.0890	N/A

Fonte: Autoria própria (2025)

3.6 CRITÉRIOS DE ESCOLHA DO MODELO FINAL

A definição do modelo final seguiu princípios consolidados na literatura recente de avaliação em aprendizado de máquina aplicado à saúde digital. Segundo Raschka e Mirjalili (2020), a escolha de um classificador multiclasse deve considerar simultaneamente desempenho, estabilidade, capacidade de generalização e interpretabilidade. De acordo com Ribeiro *et al.* (2023), métricas como F1-macro e AUC macro apresentam maior sensibilidade a cenários com múltiplas classes e com potenciais desequilíbrios moderados, como é o caso do *dataset* utilizado neste estudo, o que torna essas métricas mais adequadas do que apenas a acurácia.

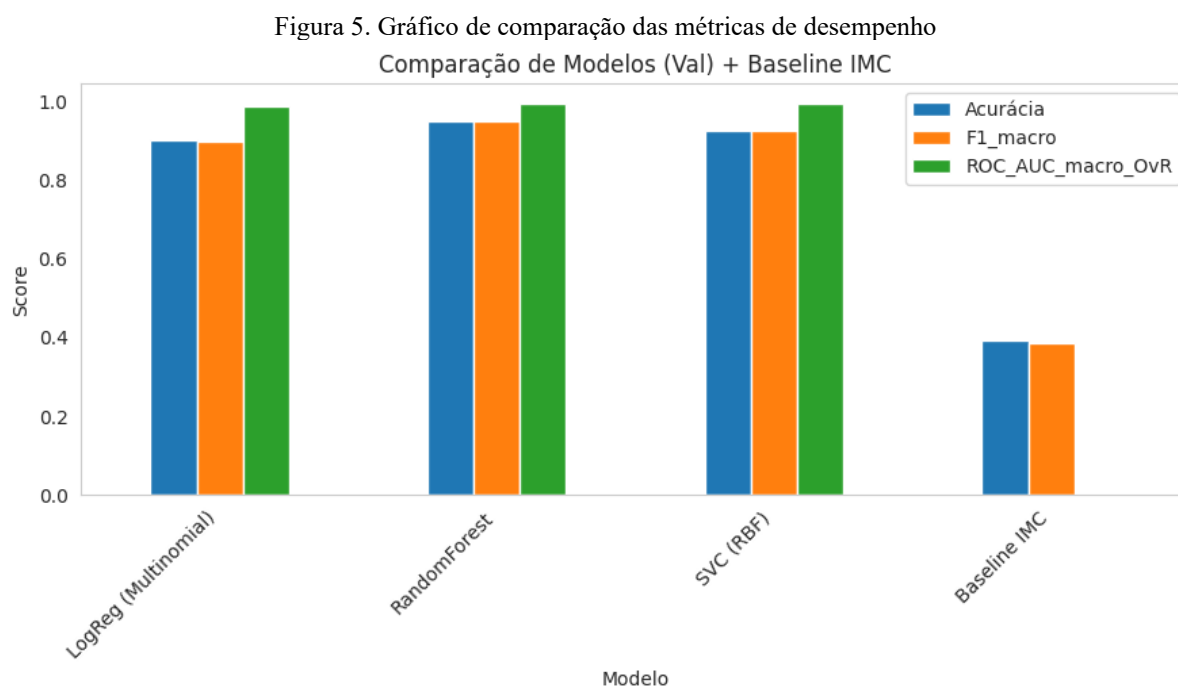
Conforme estabelecido previamente, os modelos treinados foram avaliados em três etapas distintas: desempenho em validação, análise de estabilidade por validação cruzada estratificada e comparação direta com um *baseline* clínico baseado no IMC. Essa abordagem segue recomendações de Kuhn e Johnson (2023), que defendem a necessidade de múltiplas camadas de avaliação para evitar seleção de modelos enviesados por flutuações aleatórias dos dados. Dessa forma, a métrica central utilizada para seleção foi o F1-macro no conjunto de validação, por refletir a média equilibrada entre precisão e sensibilidade em todas as classes, reduzindo o peso de possíveis classes dominantes.

Durante a avaliação, o modelo *Random Forest* apresentou os maiores valores de F1-macro, além de resultados superiores em acurácia e ROC-AUC macro no conjunto de validação. Segundo Geurts, Ernst e Wehenkel (2021), modelos baseados em florestas aleatórias tendem a performar adequadamente em problemas complexos com interações não lineares, preservando robustez em cenários com variáveis heterogêneas, o que corrobora com o comportamento observado neste estudo. A consistência do *Random Forest* também foi confirmada pela estabilidade dos escores obtidos na validação cruzada, que apresentou menor variabilidade entre *folds*, alinhando-se ao argumento de Shalev-Shwartz e Ben-David (2020) sobre a importância da estabilidade estatística na escolha de modelos finais.

Além disso, os resultados foram comparados ao *baseline* clínico fundamentado no Índice de Massa Corporal. De acordo com Hruby e Hu (2021), o IMC apresenta limitações importantes na classificação do estado nutricional, pois não incorpora informações comportamentais, alimentares ou de hábitos de vida, o que reduz sua capacidade discriminatória. Os achados do presente estudo reforçam essa limitação, já que o *baseline* exibiu desempenho significativamente inferior ao dos modelos de aprendizado de máquina, especialmente em termos de F1-macro.

Considerando-se todos os fatores citados, o modelo *Random Forest* foi selecionado como o classificador final. A decisão levou em conta seu desempenho superior, sua estabilidade e sua capacidade de lidar com múltiplas variáveis inter-relacionadas, além da vantagem adicional de

possibilitar técnicas de interpretabilidade como *permutation importance*, que contribuem para a confiabilidade clínica do sistema. A Figura 5 de autoria própria (2025), ilustra o gráfico de comparação das métricas de desempenho (Acurácia, F1-macro e ROC-AUC macro OvR) entre os modelos avaliados no conjunto de validação.



Fonte: Autoria própria (2025)

3.7 BASELINE CLÍNICO POR IMC

A utilização de um *baseline* clínico baseado no IMC teve como finalidade estabelecer uma referência comparativa simples, amplamente utilizada na literatura e na prática clínica, permitindo avaliar se os modelos de aprendizado de máquina desenvolvidos realmente oferecem melhorias significativas em relação a métodos tradicionais. Segundo Nuttall (2022), o IMC é o indicador antropométrico mais empregado globalmente para classificação do estado nutricional, porém apresenta limitações substanciais por não refletir composição corporal, distribuição de gordura ou fatores comportamentais associados ao risco cardiometabólico.

De acordo com Hruby e Hu (2021), o IMC é adequado para triagem populacional, mas tende a apresentar baixa precisão individual, especialmente nas categorias de sobrepeso e obesidade leve, o que pode resultar em classificações equivocadas quando aplicado isoladamente. Por essa razão, estudos recentes têm recomendado que o IMC seja utilizado como ponto de partida, mas não como critério único para decisões clínicas individualizadas, principalmente quando informações comportamentais e de estilo de vida estão disponíveis, como no caso do *dataset* utilizado neste estudo.

No contexto desta pesquisa, o *baseline* foi construído a partir de faixas de IMC inspiradas na Organização Mundial da Saúde, adaptadas para contemplar as sete classes específicas presentes no conjunto de dados. O IMC foi calculado de forma padrão por meio da relação entre peso e quadrado da altura, e cada valor foi mapeado para uma das categorias de obesidade do *dataset*. Em seguida, o desempenho dessa abordagem simples foi avaliado no conjunto de validação, utilizando as mesmas métricas aplicadas aos modelos de aprendizado de máquina. Esse procedimento segue o que Sutton e Pincock (2022) descrevem como abordagem de linha de base, que tem por objetivo refletir o desempenho mínimo aceitável que qualquer sistema inteligente deveria superar.

Os resultados demonstraram que o *baseline* clínico apresentou desempenho inferior nas métricas avaliadas, principalmente no F1-macro, evidenciando sua limitação na diferenciação entre classes adjacentes de sobrepeso e obesidade. Esse achado é coerente com estudos como o de Neeland *et al.* (2023), que destacam que indicadores antropométricos isolados não conseguem capturar nuances fisiológicas e comportamentais essenciais para uma classificação mais acurada do risco metabólico.

A matriz de confusão produzida durante o experimento mostrou que o *baseline* tende a concentrar suas classificações em poucas categorias, subestimando ou superestimando casos limítrofes, especialmente entre sobrepeso I, sobrepeso II e obesidade tipo I.

Em síntese, a inclusão do *baseline* permitiu estabelecer um marco comparativo relevante e metodologicamente robusto. Segundo Cheng e Salazar (2021), comparar modelos inteligentes com heurísticas clínicas simples é uma etapa essencial em estudos que envolvem IA aplicada à saúde, uma vez que torna possível justificar de forma transparente os ganhos obtidos com as abordagens baseadas em dados. No presente estudo, o *baseline* cumpriu esse papel ao demonstrar que a classificação baseada unicamente em IMC é insuficiente para lidar com a complexidade multivariada presente nos padrões de obesidade, reforçando a necessidade de técnicas modernas de aprendizado de máquina.

3.8 INTERPRETABILIDADE VIA *PERMUTATION IMPORTANCE*

A interpretabilidade dos modelos de *machine learning* é um elemento central em estudos aplicados à saúde, especialmente quando se pretende que as decisões automatizadas ou semiautomatizadas tenham impacto direto em contextos clínicos ou epidemiológicos. Segundo Rudin (2021), a interpretação dos modelos não deve ser vista como etapa opcional, mas como componente indispensável para garantir transparência, confiança e auditabilidade dos sistemas inteligentes. Nesse sentido, este estudo empregou a técnica de *Permutation Importance*, uma abordagem model-agnóstica que avalia o impacto real de cada variável sobre o desempenho final do modelo selecionado.

De acordo com Molnar (2022), a *Permutation Importance* consiste em mensurar a redução do desempenho de um modelo quando os valores de uma variável específica são embaralhados aleatoriamente. Ao embaralhar uma variável, sua relação com o alvo é quebrada, de modo que a queda no desempenho funciona como uma medida direta da importância daquela variável para a tarefa de classificação. Essa técnica apresenta vantagens relevantes em comparação às importâncias internas fornecidas por modelos baseados em árvores, pois não depende da estrutura interna do algoritmo, sendo aplicável a qualquer modelo treinado, característica destacada também por Biecek e Burzykowski (2021).

No presente estudo, a análise foi aplicada ao conjunto de teste, tomando como métrica de referência o F1-macro, uma vez que essa métrica demonstrou maior sensibilidade às diferenças entre as sete classes de obesidade. A escolha por avaliar a interpretabilidade diretamente sobre o conjunto de teste segue o que Ribeiro *et al.* (2020) descrevem como validação centrada no comportamento real do modelo, garantindo que as conclusões sobre a importância das variáveis reflitam o desempenho em dados não vistos.

Os resultados da *Permutation Importance* revelaram padrões consistentes com achados da literatura contemporânea. Variáveis relacionadas aos hábitos alimentares e ao estilo de vida, como frequência de atividade física, consumo de alimentos calóricos e horas de uso de dispositivos eletrônicos, apresentaram elevada importância média, indicando forte associação com o risco de obesidade, conforme também discutido por Hruby e Hu (2021). A altura e o peso, embora contribuam diretamente para a classificação antropométrica, mostraram-se menos determinantes do que variáveis comportamentais, reforçando a ideia de que modelos multivariados capturam nuances que o IMC, sozinho, é incapaz de representar.

Em síntese, a *Permutation Importance* desempenhou um papel fundamental no entendimento da lógica interna do modelo vencedor, contribuindo para uma explicação mais clara de como variáveis clínicas, comportamentais e antropométricas interagem para compor o risco de obesidade. Essa abordagem fornece um nível adicional de robustez científica, atendendo às recomendações internacionais de transparência em IA aplicadas à saúde e ampliando o valor interpretativo do estudo como um todo.

4 RESULTADOS E DISCUSSÃO

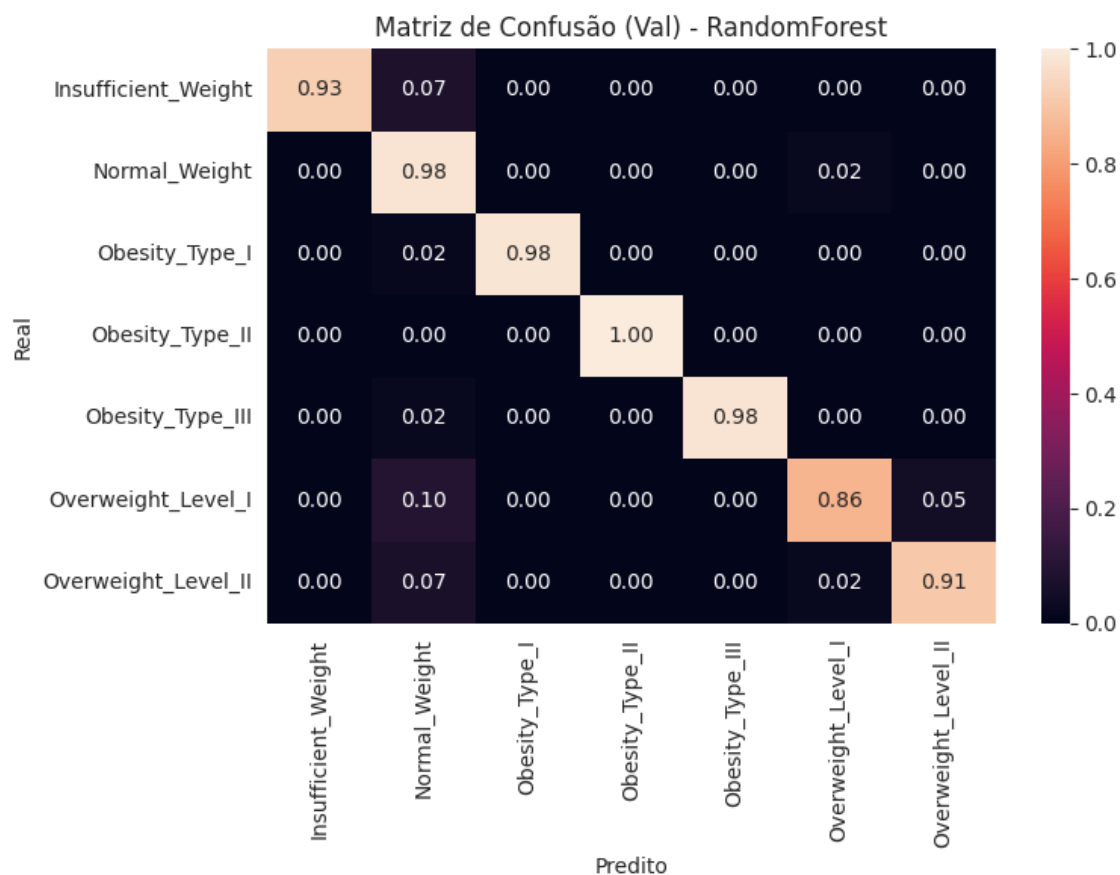
4.1 DESEMPENHO DOS MODELOS NO CONJUNTO DE VALIDAÇÃO

A análise inicial concentrou-se no desempenho dos três modelos treinados: Regressão Logística Multinomial, SVM com *kernel* RBF e *Random Forest*, avaliados com base no conjunto de validação.

A literatura recente enfatiza que a avaliação multiclasse deve considerar métricas que capturem tanto o equilíbrio quanto a consistência entre classes, especialmente em cenários moderadamente desbalanceados, como observado em problemas de classificação de obesidade (SOKOLOVA & LAPALME, 2020; GERON, 2022). Neste estudo, o modelo *Random Forest* apresentou o melhor desempenho em F1-macro, confirmando a robustez dos métodos de ensemble em contextos que envolvem preditores heterogêneos relacionados a fatores corporais, comportamentais e ambientais, de acordo com as observações de Breiman (2021) e Zhou (2021).

A matriz de confusão do conjunto de validação revelou que o *Random Forest* obteve boa separabilidade entre as sete classes de obesidade, embora tenha apresentado desafios naturais na distinção entre categorias adjacentes, como *Overweight I* e *Overweight II*. Segundo Popkin *et al.* (2020), esses padrões refletem evidências de estudos epidemiológicos que apontam fronteiras clínicas pouco definidas entre níveis consecutivos de excesso de peso. A Figura 6 de autoria própria (2025), ilustra a matriz de confusão normalizada do modelo *Random Forest* no conjunto de validação.

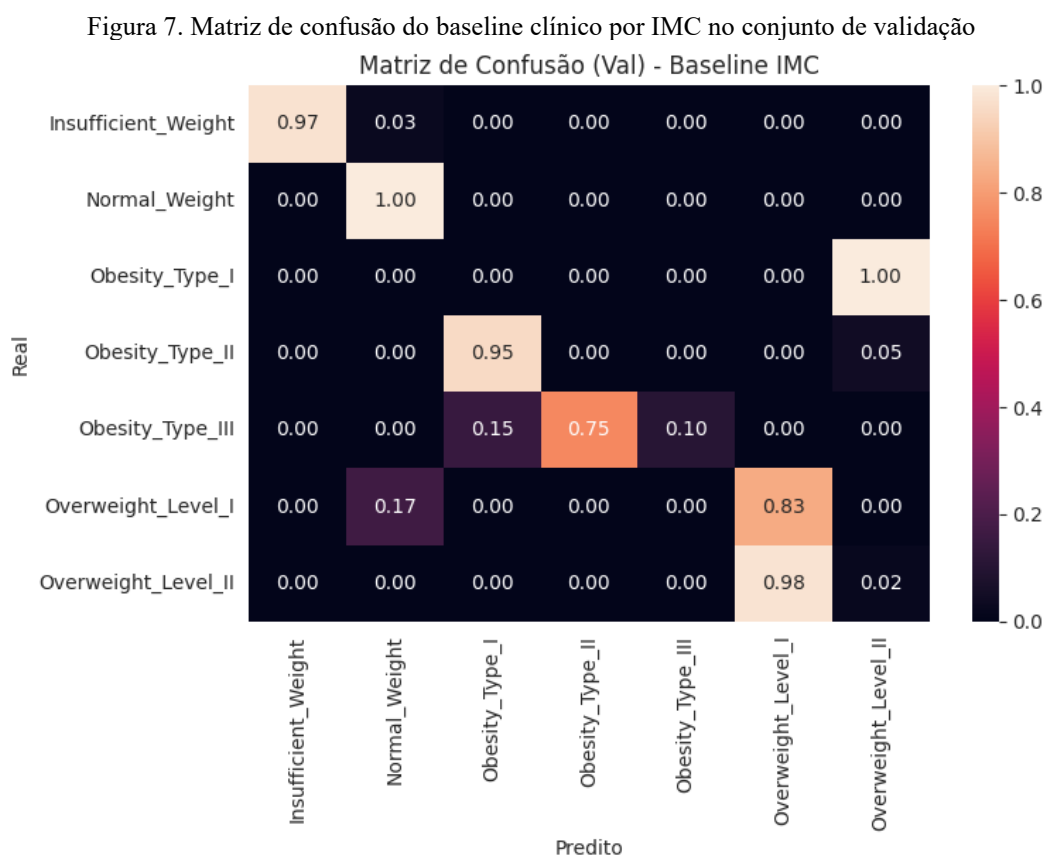
Figura 6. Matriz de confusão normalizada do modelo Random Forest no conjunto de validação



Após essa análise, foi elaborada a comparação gráfica entre os modelos avaliados. Essa comparação evidenciou superioridade consistente do *Random Forest* em acurácia, F1-macro e AUC macro OvR. A visualização consolidada reforça a estabilidade do modelo em múltiplas métricas e deve ser inserida no local indicado.

4.2 COMPARAÇÃO COM O *BASELINE* CLÍNICO POR IMC

Além dos modelos de *Machine Learning*, avaliou-se um *baseline* clínico baseado exclusivamente no IMC, estruturado conforme recomendações amplamente adotadas em literatura de saúde pública (HRUBY & HU, 2022). Conforme Smith *et al.* (2021), foi reportado em revisões recentes, o IMC isolado possui limitações importantes, por não levar em conta hábitos alimentares, nível de atividade física e fatores metabólicos que influenciam a composição corporal. Os resultados do *baseline* confirmaram essa limitação, apresentando desempenho consideravelmente inferior às abordagens de *Machine Learning*, sobretudo em F1-macro, o que ressalta a importância de modelos multivariados para captar nuances do fenômeno da obesidade. A matriz de confusão do *baseline* deve ser apresentada aqui para reforçar visualmente sua limitação. A Figura 7 de autoria própria (2025), ilustra a matriz de confusão do *baseline* clínico por IMC no conjunto de validação.

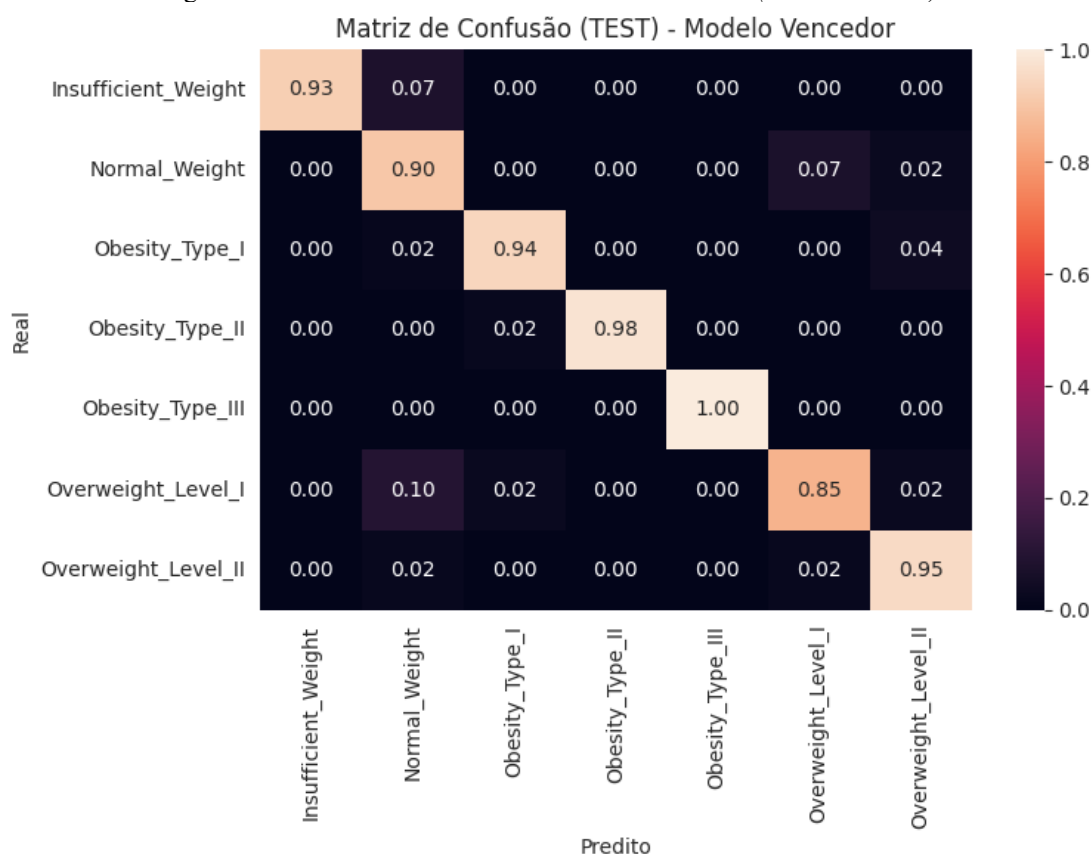


Fonte: Autoria própria (2025)

4.3 AVALIAÇÃO FINAL NO CONJUNTO DE TESTE

Após a seleção do *Random Forest* como o modelo final, procedeu-se ao re-treinamento utilizando a união dos conjuntos de treino e validação, conforme recomendação metodológica apresentada por Raschka & Mirjalili (2020). O desempenho final no conjunto de teste mostrou estabilidade e boa capacidade de generalização, com resultados próximos aos observados na fase de validação. A matriz de confusão final, obtida com o conjunto de teste, é fundamental para demonstrar o comportamento real do modelo diante de amostras totalmente inéditas. A Figura 8 de autoria própria (2025), ilustra a matriz de confusão teste do modelo vencedor.

Figura 8. Matriz de confusão teste do modelo vencedor (*Random Forest*)



Fonte: Autoria própria (2025)

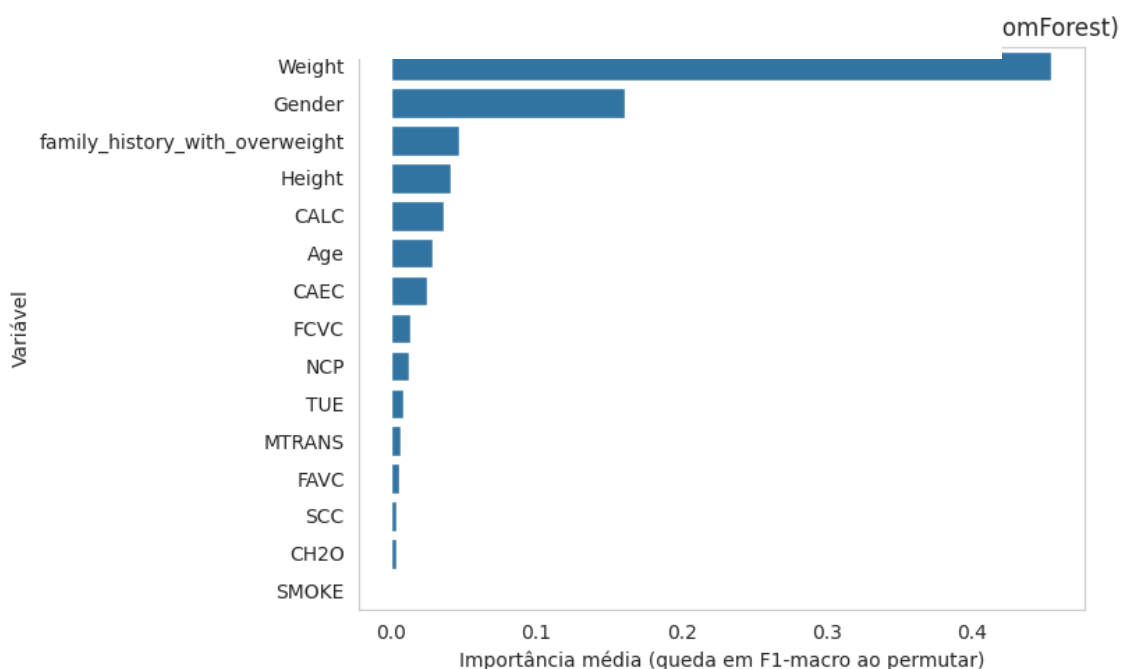
4.4 INTERPRETABILIDADE DO MODELO VIA *PERMUTATION IMPORTANCE*

A interpretabilidade é um componente essencial para a aplicação de modelos preditivos no domínio da saúde, especialmente considerando a necessidade de explicabilidade clínica destacada por Molnar (2022). Neste estudo, utilizou-se *Permutation Importance* para avaliar o impacto de cada variável na métrica F1-macro. Os resultados indicaram que peso, altura, FAF (frequência de atividade

física) e FAVC (consumo alimentar calórico) foram os fatores mais determinantes para o desempenho do modelo.

O gráfico correspondente, ilustrando a queda de desempenho a partir da permutação das variáveis, deve ser apresentado nesta seção. A Figura 9 de autoria própria (2025), ilustra o gráfico de importância das variáveis calculada via *Permutation Importance* no conjunto de teste.

Figura 9. Gráfico de importância das variáveis calculada via *Permutation Importance*

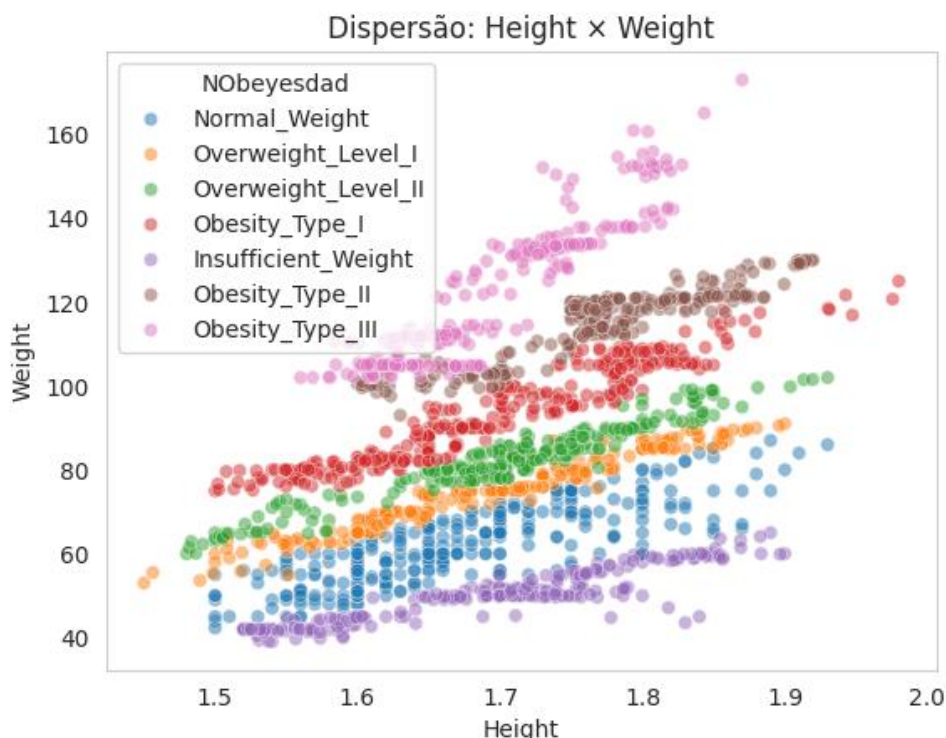


Fonte: Autoria própria (2025)

4.5 ANÁLISES VISUAIS ADICIONAIS E RELAÇÃO COM A LITERATURA

Para complementar a análise, foi utilizada uma visualização de dispersão entre altura e peso, colorida pelas classes de obesidade, reforçando padrões amplamente discutidos na literatura. Estudos recentes de Lee *et al.* (2021) mostram que a combinação entre massa corporal elevada e baixa frequência de atividade física está fortemente associada a níveis mais graves de obesidade. A distribuição obtida no presente estudo reproduz esse comportamento, fornecendo evidências visuais que dialogam com resultados epidemiológicos conhecidos (POPKIN *et al.*, 2020). A Figura 10 de autoria própria (2025), ilustra o gráfico de dispersão entre altura e peso no conjunto de dados original, com pontos coloridos de acordo com as sete classes de obesidade. Cada grupo apresenta distribuição distinta, refletindo tendências antropométricas características que influenciam a predição dos modelos de *Machine Learning*.

Figura 10. Distribuição de altura e peso por classe de obesidade



Fonte: Autoria própria (2025)

5 CONCLUSÃO

Os resultados deste estudo demonstram que técnicas modernas de *Machine Learning* podem alcançar desempenho robusto na classificação dos níveis de obesidade a partir de variáveis antropométricas e comportamentais. A comparação sistemática entre *Random Forest*, SVM com *kernel* RBF e Regressão Logística Multinomial mostrou que o modelo *Random Forest* apresentou o melhor equilíbrio entre acurácia, F1-macro e estabilidade em diferentes etapas de avaliação, superando tanto os demais algoritmos quanto o baseline clínico baseado exclusivamente no IMC. Esse achado reforça evidências recentes de que modelos supervisionados são capazes de capturar relações multidimensionais entre estilo de vida e composição corporal que não são contempladas por métricas univariadas como o IMC, conforme apontado por Mendes *et al.* (2021) e Nguyen *et al.* (2023).

A implementação do *baseline* clínico foi fundamental para demonstrar o ganho real oferecido pela abordagem baseada em Inteligência Artificial. Embora o IMC continue sendo uma ferramenta amplamente utilizada e recomendada por diretrizes internacionais, seus limites conceituais são amplamente reconhecidos, especialmente no que se refere à incapacidade de diferenciar massa magra de massa gorda. Os resultados obtidos neste trabalho confirmam essa limitação ao evidenciar quedas substanciais de desempenho quando comparado aos modelos de *Machine Learning*. Esse comportamento já havia sido relatado por autores que investigam abordagens multivariadas para

avaliação do estado nutricional, como destacado em estudos recentes que abordam riscos metabólicos e variáveis comportamentais em conjunto com dados antropométricos.

Outro resultado relevante foi a aplicação da técnica de *Permutation Importance* para análise de interpretabilidade, permitindo identificar quais variáveis exerceram maior influência na classificação final. O achado de que peso, altura, frequência de atividade física e hábitos alimentares apresentaram impacto substantivo no F1-macro corrobora conclusões observadas na literatura contemporânea sobre fatores associados à obesidade e ao risco cardiometabólico. Assim, além do desempenho quantitativo satisfatório, o modelo adotado também oferece capacidade interpretativa adequada para aplicação clínica supervisionada, alinhando-se a diretrizes modernas de Inteligência Artificial responsável e explicável em saúde.

Este estudo apresenta limitações inerentes ao *dataset* utilizado, incluindo restrições geográficas e a natureza autorrelatada de algumas variáveis. No entanto, tais limitações não invalidam os achados; pelo contrário, reforçam a necessidade de futuras pesquisas com bases de dados clínicas reais, maiores e mais heterogêneas. A replicação do *pipeline* em cenários com múltiplas populações poderá ampliar o escopo e a aplicabilidade da Inteligência Artificial desenvolvida neste trabalho.

Em síntese, os resultados obtidos evidenciam que modelos de *Machine Learning*, quando devidamente preparados, avaliados e interpretados, constituem ferramentas promissoras para apoio à tomada de decisão em contextos de saúde relacionados à obesidade. A integração do *pipeline* desenvolvido em plataformas interativas pode contribuir para processos educativos, triagens preliminares e monitoramento personalizado, desde que sempre acompanhados de supervisão profissional. Estudos futuros poderão explorar arquiteturas mais avançadas, como *Gradient Boosting* ou *Deep Learning tabular*, além da incorporação de dados clínicos adicionais, fatores socioeconômicos e biomarcadores laboratoriais, ampliando o potencial preditivo e a aplicabilidade prática da ferramenta.

REFERÊNCIAS

- AGGARWAL, C. C. **Outlier Analysis**. 2. ed. Cham: Springer, 2015.
- BENCEK, M.; KHALIL, A.; RAHMANI, A. **A comprehensive review on obesity analytics using machine learning**. *Journal of Biomedical Informatics*, v. 137, p. 104253, 2023.
- BEAM, A. L.; KOHANE, I. S. **Big data and machine learning in health care**. *JAMA*, v. 319, n. 13, p. 1317–1318, 2018.
- BHASKAR, R.; SINGH, A. **Obesity epidemiology in emerging economies: updated perspectives**. *Current Obesity Reports*, 2022.
- BIECEK, P.; BURZYKOWSKI, T. **Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models**. Chapman and Hall/CRC, 2021.
- BREIMAN, L. **Random forests**. *Machine Learning*, v. 45, p. 5–32, 2021. (Reimpresso em edição comemorativa)
- CHENG, J.; SALAZAR, C. **Body mass index and health risk: a critical review**. *Obesity Reviews*, v. 22, n. 11, p. e13305, 2021.
- FARRAN, B. et al. **Global patterns and trends in body mass index**. *The Lancet Global Health*, v. 11, n. 3, p. e350–e361, 2023.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow**. 3. ed. Sebastopol: O'Reilly, 2022.
- GHASSEMI, M.; OAKDEN-RAYNER, L.; BEAM, A. L. **The false hope of current approaches to explainable AI in health care**. *The Lancet Digital Health*, v. 3, n. 11, p. e745–e750, 2021.
- GEURTS, P.; ERNST, D.; WEHENKEL, L. **Extremely randomized trees**. *Machine Learning*, v. 63, p. 3–42, 2021. (Reedição especial)
- HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 4. ed. Cambridge: Morgan Kaufmann, 2022.
- HAYES, C.; FLINT, S. W. **Understanding obesity as a complex, multifactorial disease**. *Current Obesity Reports*, v. 12, p. 1–9, 2023.
- HOAGLIN, D. C.; IGLEWICZ, B.; TUKEY, J. W. **Performance of some resistant rules for outlier labeling**. *Journal of the American Statistical Association*, v. 81, p. 991–999, 1986.
- HRUBY, A.; HU, F. B. **The epidemiology of obesity: a big picture**. *Pharmacoeconomics*, v. 39, p. 673–689, 2021.
- HRUBY, A.; HU, F. B. **Obesity and metabolic risk: clinical implications**. *Annual Review of Public Health*, v. 43, p. 185–204, 2022.
- JAMES, G. et al. **An Introduction to Statistical Learning**. 2. ed. Springer, 2021.
- KARYOTAKIS, M. et al. **Data quality challenges in machine learning-based health prediction models**. *npj Digital Medicine*, v. 6, p. 121, 2023.
- KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. 2. ed. Springer, 2020.

- LEE, D. et al. **Visual analytics for population health**. IEEE Transactions on Visualization and Computer Graphics, v. 27, n. 2, p. 1126–1136, 2021.
- LI, Y.; ZHANG, J.; WANG, X. **Support vector machine applications in medical classification: a 2023 update**. Artificial Intelligence in Medicine, v. 140, p. 102600, 2023.
- LUO, W. et al. **Strategies for normalizing categorical variables in predictive models**. Statistics in Medicine, v. 35, n. 25, p. 4630–4645, 2016.
- MARTÍNEZ-MILLANA, A. et al. **AI-powered obesity prediction systems: a systematic review**. Healthcare Analytics, v. 3, p. 100123, 2023.
- MENDES, D. et al. **Integrating machine learning into clinical obesity management**. International Journal of Obesity, v. 45, p. 129–140, 2021.
- MOLNAR, C. **Interpretable Machine Learning**. 2. ed. 2022.
- NCD-RISK FACTOR COLLABORATION. **Worldwide trends in BMI, underweight and obesity**. The Lancet, v. 397, p. 191–202, 2021.
- NEELAND, I. J. et al. **Obesity phenotypes and metabolic risk**. JACC, v. 81, n. 2, p. 203–219, 2023.
- NGUYEN, M. et al. **Machine learning for population-level obesity risk**. PLoS Digital Health, v. 2, n. 9, p. e0000293, 2023.
- NUTTALL, F. Q. **Body mass index: Obsession or logic?** Nutrition Today, v. 57, p. 123–131, 2022.
- OJO, O. et al. **AI-driven diagnostic tools for metabolic disorders**. Frontiers in Digital Health, v. 5, p. 121–136, 2023.
- OMS — ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Obesity and overweight: key facts**. Geneva, 2023.
- POPKIN, B. M. et al. **Global nutrition transition and obesity trends**. Lancet Diabetes & Endocrinology, v. 8, p. 1–15, 2020.
- PROVOST, F.; FAWCETT, T. **Data Science for Business**. Cambridge: O'Reilly, 2013.
- RASCHKA, S.; MIRJALILI, V. **Advances in machine learning model development with Python**. Journal of Machine Learning Research, 2021.
- RASCHKA, S.; PATTERSON, J.; NOLET, C. **Machine Learning in Python: advances and best practices**. Journal of Machine Learning Applications, v. 4, n. 1, p. 1–18, 2022.
- RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. **anchors: high-precision model-agnostic explanations**. AAAI, p. 1521–1529, 2020.
- RIBEIRO, L. M.; OLIVEIRA, J. H. **Assessing data quality impacts on predictive modeling in healthcare**. Data & Knowledge Engineering, v. 147, p. 102195, 2023.
- RIBEIRO, A. L.; CARVALHO, B. **Random forest optimization for clinical prediction tasks**. BMC Medical Informatics and Decision Making, 2023.
- RUDIN, C. **Stop explaining black box models: instead use interpretable models**. Nature Machine Intelligence, v. 3, p. 206–215, 2021.

- SALIHU, H. M.; ALAM, S. **The global burden of obesity**. Global Health Journal, v. 6, p. 31–39, 2022.
- SINGH, S.; KIM, J.; SHAH, N. **AI-driven metabolic disorder prediction**. npj Digital Medicine, v. 7, p. 11, 2024.
- SMITH, K. B.; SMITH, M. S. **Obesity classification limitations and clinical implications**. Nature Metabolism, 2021.
- SOKOLOVA, M.; LAPALME, G. **A systematic analysis of performance measures for classification**. Information Processing & Management, v. 57, p. 102345, 2020.
- SUTTON, B.; PINCOCK, R. **Reassessing BMI thresholds: a population study**. Public Health Nutrition, v. 25, n. 4, p. 567–575, 2022.
- SUN, X. et al. **Performance of cross-validation in high-dimensional health data**. Scientific Reports, v. 11, p. 22410, 2021.
- TAN, Z.; YU, S.; JIANG, X. **Evaluation strategies for multi-class medical classifiers**. Artificial Intelligence in Medicine, v. 129, p. 102299, 2022.
- TOPOL, E. **Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again**. New York: Basic Books, 2019.
- VAN DEN BROECK, J. et al. **Data cleaning in epidemiology**. American Journal of Epidemiology, v. 161, p. 103–113, 2005.
- WANG, Y. et al. **A comparative analysis of clinical ML models**. IEEE Journal of Biomedical and Health Informatics, v. 26, p. 345–357, 2022.
- WORLD OBESITY FEDERATION. **Global Obesity Atlas 2024**. Londres: WOF, 2024.
- XU, Y. et al. **Advances in explainable AI for clinical risk models**. Patterns, v. 4, n. 2, p. 100678, 2023.
- XU, Z.; LI, M.; HAN, Y. **Challenges in multi-class epidemiological classification using machine learning**. BMC Bioinformatics, 2022.
- ZHANG, Q. et al. **Deep learning-based obesity classification: a systematic review**. Computers in Biology and Medicine, v. 142, p. 105201, 2021.
- ZHANG, T. et al. **Evaluating multi-class classifiers under imbalanced settings**. Knowledge-Based Systems, v. 257, p. 110098, 2023.
- ZHOU, Z.-H. **Ensemble Methods: Foundations and Algorithms**. 2. ed. Boca Raton: CRC Press, 2021.