



UnB

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE MEDICINA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS MÉDICAS

JÉSSICA APARECIDA SILVA

AVALIAÇÃO DO MODELO DE APRENDIZADO DE MÁQUINA COM A MELHOR PERFORMANCE PARA PREDIZER A RESISTÊNCIA INSULÍNICA EM ADOLESCENTES DO ESTUDO DE RISCOS CARDIOVASCULARES EM ADOLESCENTES.

Brasília - DF

2024

JÉSSICA APARECIDA SILVA

AVALIAÇÃO DO MODELO DE APRENDIZADO DE MÁQUINA COM A MELHOR PERFORMANCE PARA PREDIZER A RESISTÊNCIA INSULÍNICA EM ADOLESCENTES DO ESTUDO DE RISCOS CARDIOVASCULARES EM ADOLESCENTES.

Dissertação apresentada como requisito parcial para obtenção do Título de Mestre em Ciências Médicas pelo Programa de Pós-Graduação em Ciências Médicas da Universidade de Brasília.

Orientador: Prof. Dr. Rodolfo Rêgo Deusdará Rodrigues

Brasília - DF

2024

AGRADECIMENTOS

Ao professor Rodolfo Rêgo Deusdará, agradeço pela oportunidade e por todo o compartilhamento dos seus conhecimentos durante este período. Sua competência, segurança e parceria estiveram presentes em cada etapa do desenvolvimento deste estudo e serão minha inspiração acadêmica.

Um agradecimento em especial ao professor Moyses Szklo e as professoras Marília Miranda, Lucia Freitas e Katia Bloch. Muito obrigada pelas contribuições e compartilhamento de conhecimento em nossos encontros.

Com muito carinho, meus agradecimentos à Natalia e à Mirian, assistentes que me acolheram com grande paciência sempre que necessário.

Aos professores do PPGCM com os quais tive contato e colaboraram para meu crescimento, em especial meu agradecimento ao professor Daniel Barroso, pela parceria durante a mentoria da disciplina de métodos quantitativos.

Agradeço à equipe ERICA pela disponibilização do Banco de Dados. À UnB por proporcionar as condições para um aprendizado de excelência.

Um agradecimento repleto de amor à minha família. A minha esposa Isadora pela constante torcida, incentivo e amor nesta decisão de continuar a uma trajetória de estudos. A sua presença é preciosa em minha vida.

Eternamente agradeço aos meus pais, que embora hoje não participem ativamente deste processo, sempre serão minha inspiração e terão importância primordial em minha vida. Enfim, a todos que direta ou indiretamente torceram por este trabalho, especialmente à DEUS por suas bençãos.

RESUMO

Introdução: A resistência à insulina é definida como uma diminuição da resposta dos tecidos às ações celulares mediadas pela insulina, sendo o inverso da sensibilidade à insulina. Atualmente existem métodos considerados padrão-ouro para medir a resistência à insulina, como o clamp euglicêmico hiperinsulinêmico e o clamp hiperglicêmico. Mas, estes métodos são caros, trabalhosos e raramente utilizados em grandes estudos epidemiológicos. **Objetivo:** Avaliar o melhor modelo preditivo de aprendizado de máquina para prever a resistência à insulina no Estudo de Riscos Cardiovasculares em Adolescentes (ERICA) 2013-2014. **Métodos:** Estudamos aproximadamente 38.000 adolescentes, utilizando as variáveis HDL colesterol, LDL colesterol, triglicerídeos, circunferência da cintura, estado nutricional, inatividade física, tabagismo, consumo de álcool, comportamento sedentário, nível socioeconômico, pressão arterial e idade para criação dos modelos de aprendizado de máquina. A amostra foi estratificada por sexo e desenvolvemos os modelos: Regressão Logística (RL), *Poisson*, Rede Neural Profunda (RNP), XGBoost, *Random Forest* (RF), *Decision Tree* (DT) e *Support Vector Machine* (SVM). Avaliámos os modelos utilizando a área sob a curva (AUC), as curvas de calibração e a análise da curva de decisão (DCA) e o gráfico *SHapley Additive exPlanations*. **Resultados:** Os modelos RL, *Poisson*, XGBoost, RNP e RF apresentaram a AUC mais elevada, variando entre 0,72 e 0,8 para meninos e meninas. Três modelos (RL, XGBoost e RNP) apresentaram uma calibração mais elevada nas meninas do que nos meninos, 80% e 60%, respectivamente. A curva DCA mostrou valores de prevalência em torno de 18% para as meninas e 16% para os meninos e, para além deste limiar de probabilidade, todos os modelos, com exceção do modelo SVM, poderiam ser benéficos na intervenção clínica. Por fim, os preditores clínicos mais importantes encontrados para ambos os sexos foram a circunferência da cintura, triglicerídeos e idade. **Conclusões:** Nosso estudo sugere que a RL pode ser útil para prever a resistência à insulina em adolescentes do ERICA, pois destaca-se como a melhor escolha devido à sua interpretabilidade, ampla aplicabilidade em contextos clínicos, eficiência computacional e desempenho comparável aos outros modelos desenvolvidos.

Palavras-chave: aprendizado de máquina, resistência insulínica, adolescentes, saúde pública, HOMA-IR.

ABSTRACT

Introduction: Insulin resistance is defined as a decreased tissue response to insulin-mediated cellular actions, and it is the inverse of insulin sensitivity. There are currently methods considered the gold standard for measuring insulin resistance, such as the hyperinsulinemic euglycemic clamp and the hyperglycemic clamp. However, these methods are expensive and are rarely used in large epidemiological studies. **Aim:** Assessment of machine learning models performance for clinical prediction of insulin resistance in The Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. **Methods:** We studied approximately 38,000 adolescents and considered HDL cholesterol, LDL cholesterol, triglycerides, waist circumference, nutritional status, physical inactivity, smoking habits, alcohol consumption, sedentary behavior, socioeconomic status, blood pressure and age variables. We began stratifying the subset by sex and create the model building process for machine learning models logistic regression (LR), Poisson, Deep Neural Network (DNN), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM). We evaluate the models using area under the curve (AUC), calibration curves and decision curve analysis (DCA). Finally, we used SHapley Additive exPlanations. **Results:** LR, Poisson, XGBoost, DNN and RF models had the highest AUC ranging from 0.72 to 0.8 for boys and girls. Three models (LR, XGBoost and DNN) had higher calibration in girls than in boys, up to 80% vs up to 60%, respectively. The DCA curve showed prevalence values around 18% for girls and 16% for boys, and beyond this threshold of probability, all models could provide beneficial in clinical intervention, with exception of SVM model. Finally, the most important clinical predictors for both sexes were waist circumference, triglycerides and age. **Conclusions:** Our study suggests that the LR model could be useful to predict insulin resistance in ERICA adolescents and stands out as the most appropriate choice in this context due to its interpretability, broad applicability in clinical settings, computational efficiency, and comparable performance to other developed models.

Keywords: machine learning, insulin resistance, adolescents, public health, HOMA-IR.

LISTA DE FIGURAS

Figura 1. Desenho esquemático do algoritmo <i>decision tree</i> contendo 8 nós e 7 ramos.....	23
Figura 2. Desenho esquemático do modelo <i>random forest</i> demonstrando a presença de diversas árvores de decisão.....	24
Figura 3. Estrutura geral do algoritmo SVM. O algoritmo define o hiperplano de separação entre classes, onde a seleção desses vetores de suporte visam maximizar a margem de separação entre as classes.....	25
Figura 4. Camadas de uma Rede Neural e interação entre os neurônios. A rede neural mostrada é formada por três camadas sendo elas a de entrada, oculta e saída.....	26
Figura 5. Exemplo de um modelo XGBoost contendo 3 árvores que buscam minimizar a função de perda para obtenção do melhor resultado.....	27
Figura 6. Exemplo de um modelo Catboost contendo 3 árvores que buscam minimizar a função de perda para obtenção do melhor resultado.....	28
Figura 7. Etapas do fluxo de Aprendizado de Máquina desenvolvido para treinar modelos de aprendizado de máquina com o objetivo de prever a resistência à insulina no Estudo de Riscos Cardiovasculares em Adolescentes (ERICA) 2013-2014. Dos 37.815 participantes potencialmente elegíveis para o desenvolvimento do modelo, 37.454 adolescentes atenderam aos critérios de inclusão.....	40

Manuscrito

Figure 1. Machine Learning Pipeline developed to train machine learning models to predict insulin resistance in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. Out of 37.815 potentially eligible participants for model development, 37.454 adolescents met the inclusion criteria, consisting of students from the morning shift.....	48
Figure 2. Distribution of the Area Under the Curve (AUC) for each machine learning algorithm developed to predict insulin resistance in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. The AUC distributions were calculated using the	

test dataset.....	50
Figure 3. The calibration curve analysis for each machine learning algorithm developed to predict insulin resistance, in the test dataset, in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014.....	52
Figure 4. The decision curve analysis for each machine learning algorithm developed to predict insulin resistance in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. Each curve was calculated using the training dataset.....	53
Figure 5. The importance of variables according to the Shapley values for the Logistic Regression model on the test dataset of the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. The vertical axis lists the input variables, and the horizontal axis measures their contribution to classifying cognitive deficits. The red color represents higher values, and blue represents lower values. SHAP: Shapley Additive explanations.....	53

LISTA DE TABELAS

Manuscrito

Table 1. Characteristics of 37,815 adolescents enrolled in the Study of Cardiovascular Risk in Adolescents (ERICA, 2013–2014).....	49
Table 2. Performance metrics of machine learning models in predicting insulin resistance according to sex in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014.....	51
Supplementary Table 1. Sensitivity analysis comparing simple survey design, complex survey design and the the 75th Percentile Value as the Cut-Off Value for Insulin Resitance (HOMA-IR). Metrics calculated to evaluate the performance of machine learning models in predicting insulin resistance in the Study of Cardiovascular Risk in Adolescents (ERICA) 2013-2014. The metrics were calculated using the training dataset of girls and boys.....	62

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de máquina
AUC	<i>Area Under the Curve</i>
aOR	<i>Adjusted Odds Ratios</i>
CatBoost	<i>Categorical Boosting</i>
DCV	Doenças Cardiovasculares
DCA	<i>Decision Curve Analysis</i>
DM2	Diabetes Mellitus Tipo 2
DT	<i>Decision Tree</i>
ERICA	Estudo de Riscos Cardiovasculares Em Adolescentes
FIRI	<i>Fasting Insulin Resistance Index</i>
HDL Cholesterol	<i>High Density Lipoprotein Cholesterol</i>
HOMA-%B	<i>Homeostasis Model Assessment β-Cell Function</i>
HOMA-IR	<i>Homeostasis Model Assessment of Insulin Resistance</i>
IA	Inteligência Artificial
IC 95%	Intervalo de confiança de 95%
IMC	Índice de Massa Corporal
IQR	Intervalos Interquartílicos
KNN	<i>K-Nearest Neighbors</i>
LDL Cholesterol	<i>Low Density Lipoprotein Cholesterol</i>
LightGBM	<i>Light Gradient Boosting Machine</i>
MLG	Modelos Lineares Generalizados
ML	<i>Machine Learning</i>
PA	Pressão arterial
PeNSE	Pesquisa Nacional de Saúde Escolar 2019
QUICKI	<i>Quantitative Insulin Sensitivity Check Index</i>
RF	<i>Random Forest</i>
RI	Resistência insulínica
RNP	Rede Neural Profunda

RL	Regressão Logística
SHAP	<i>Shapley Additive ExPlanations</i>
SM	Síndrome Metabólica
SVM	<i>Support Vector Machine</i>
Tg/HDL	Razão de Triglicérides/HDL
TyG	Índice de Triglicérides/glicose
VPP	Valor Preditivo Positivo
VPN	Valor Preditivo Negativo
XGBoost	<i>Extreme Gradient Boosting</i>

SUMÁRIO

1. INTRODUÇÃO	13
2. REVISÃO DE LITERATURA	15
2.1 Implicações clínicas da Resistência Insulínica	15
2.2 Sensibilidade e Resistência Insulínica na Puberdade e Adolescência	15
2.3 Obesidade na Puberdade e Adolescência	16
2.4 Aspectos de Estilo de Vida: Tabagismo, Álcool, Atividade Física e Comportamento Sedentário	17
2.5 Métodos de detecção de Resistência Insulínica	19
2.6 Modelos de Aprendizado de Máquina	20
2.6.1 Modelagem preditiva de dados clínicos.....	20
2.6.2 Desenvolvimento, avaliação e comparação de modelos preditivos	21
2.6.3 Principais algoritmos de aprendizagem de máquina em predição clínica	22
2.6.3.1 Modelos lineares generalizados (MLG).....	22
2.6.3.2 Decision Tree	22
2.6.3.3 Random Forest.....	23
2.6.3.4 Support Vector Machine (SVM).....	23
2.6.3.5 Redes Neurais Profundas	24
2.6.3.6 Extreme gradient boosting (XGBoost).....	25
2.6.3.7 Categorical Boosting (CatBoost)	26
2.7 Modelos de predição de resistência insulínica	27
2.8 Preditores associados a resistência insulínica	28
3. JUSTIFICATIVA	29
4. HIPÓTESE	30
5. OBJETIVO	31
5.1 Objetivo geral	31
5.2 Objetivos específicos	31
6. METODOLOGIA	32
6.1 Estudo de Riscos Cardiovasculares em Adolescentes	32
6.1.1 Desenho amostral.....	32
6.1.2 Questionários.....	33
6.1.3 Recordatário e exames bioquímicos.....	33
6.1.4 Medidas antropométricas e pressão arterial	33
6.2 Conjunto de dados e seleção de variáveis	33

6.2.1 Colesterol HDL	33
6.2.2 Colesterol LDL.....	34
6.2.3 Triglicérides.....	34
6.2.4 Circunferência da cintura.....	34
6.2.5 Inatividade Física	34
6.2.6 Hábitos de fumo	34
6.2.7 Consumo de álcool.....	35
6.2.8 Comportamento sedentário	35
6.2.9 Variável socioeconômica.....	35
6.2.10 Estado nutricional	35
6.2.11 Pressão Arterial.....	35
6.2.12 Modelo Homeostático de Avaliação da Resistência à Insulina.....	36
6.3 Desenho do estudo	36
6.4 Desenvolvimento e avaliação dos modelos de aprendizado de máquina....	38
6.5 Aspectos éticos	39
7. RESULTADOS.....	40
7.1 Manuscrito	40
8. CONSIDERAÇÃO FINAIS.....	41
REFERÊNCIAS.....	42

1. INTRODUÇÃO

Segundo um boletim das Nações Unidas, cerca de 21% da população mundial são adolescentes com idade entre 10 e 19 anos (UNITED NATIONS, 2019; WORLD HEALTH ORGANIZATION, 2020). A adolescência é considerada uma fase de transição entre a infância e a idade adulta, durante a qual ocorrem mudanças físicas, psicológicas e sociais intensas. Geralmente começa com a puberdade e termina por volta dos 19 anos (TANNER, 1962).

A resistência insulínica (RI) pode ser definida como uma diminuição da resposta do tecido às ações celulares mediadas pela insulina e é o inverso da sensibilidade à insulina (LEVY-MARCHAL et al., 2010). Na puberdade, está frequentemente associada à ocorrência de um status de RI com aumento compensatório dos níveis de insulina (GORAN et al., 2001).

Uma revisão sistemática mostrou que a prevalência de RI encontrada em adolescentes foi de 3,1% até 44%. Essa variação pode ser explicada pelo uso de diferentes equações para definir a resistência insulínica (VAN DER AA et al., 2015). Atualmente não há uma expressão numérica universalmente aceita para definir a RI, e não há um ponto de corte claro para distinguir entre normal e anormal (LEVY-MARCHAL et al., 2010).

Para medir a RI existem métodos considerados padrão-ouro como clamp euglicêmico hiperinsulinêmico e clamp hiperglicêmico. Porém, tais métodos são caros, trabalhosos e por isso são raramente utilizados em grandes estudos epidemiológicos (SINGH et al., 2010). Em crianças e adolescentes, métodos alternativos indiretos baseados em níveis de insulina e glicose em jejum são utilizados, como o *Homeostasis Model Assesment Insulin Resistance* (HOMA-IR) (SCHWARTZ et al., 2008).

O modelo HOMA é uma ferramenta matemática que estima a resistência insulínica (HOMA-IR) e a função das células beta (HOMA-% β) com base nas concentrações de glicose e insulina no sangue em jejum. Esse modelo é composto por equações empíricas não lineares que descrevem a regulação da glicose pelos órgãos e tecidos envolvidos (MATTHEWS et al., 1985). Uma das principais vantagens de se utilizar o HOMA-IR é que para o seu cálculo não são necessárias várias

amostras de sangue, tornando-o prático em estudos epidemiológicos e clínicos (SCHWARTZ et al., 2008).

O Estudo de Riscos Cardiovasculares em Adolescentes (ERICA) é um estudo multicêntrico brasileiro, cujo objetivo é fornecer estimativas de fatores de risco cardiovascular, síndrome metabólica (SM) e seus componentes. No estudo foram avaliadas as condições de saúde de cerca de 75 mil estudantes entre 12 e 17 anos, de 1.247 escolas brasileiras, públicas e particulares, distribuídas por 122 cidades participantes (DA SILVA et al., 2016).

O estudo realizado por Chissini et al. (2020) analisou o ERICA e mostrou que os valores de corte ideais do HOMA-IR associados à síndrome metabólica na população total foram de 2,8 para os adolescentes do sexo masculino e de 2,32 para as adolescentes do sexo feminino. Esse estudo também mostrou que a resistência insulínica foi prevalente em 19,1% dos adolescentes (IC 95%), sendo a maior prevalência entre meninas com excesso de peso.

A inteligência artificial (IA) têm sido aplicada de forma eficaz no diagnóstico de doenças, tratamento personalizado e prognóstico (YU et al., 2018; OIKONOMOU et al., 2023). Pesquisas no campo de aprendizado de máquina foram desenvolvidas buscando realizar a predição de resistência insulínica em suas respectivas populações alvo. Porém, podemos notar que existem poucos estudos focando em populações de países de renda média baixa, especialmente estudos com os adolescentes como população alvo (YU et al., 2018; OIKONOMOU et al., 2023; ZHANG et al., 2022; ZHANG et al., 2024; PARK et al., 2022; LEE et al., 2022)

Buscando auxiliar profissionais de saúde na identificação da resistência insulínica, podemos utilizar modelos preditivos criados a partir de banco de dados de estudos epidemiológicos. Com isso, o objetivo do presente estudo foi desenvolver o melhor modelo preditivo de aprendizado de máquina para resistência à insulina, definido pelo HOMA-IR, baseado em dados do Estudo de Riscos Cardiovasculares em Adolescentes (BLOCH et al., 2015).

2. REVISÃO DE LITERATURA

2.1 Implicações clínicas da Resistência Insulínica

A história da resistência insulínica remonta a década de 1960, quando pesquisadores começaram a identificar uma diminuição na capacidade das células-alvo de responderem à ação da insulina, com isso demonstrou-se que indivíduos com diabetes mellitus apresentavam alta concentração de insulina em comparação com indivíduos não diabéticos (YALOW; BERSON, 1960a, 1960b). Desse modo, a RI foi definida como um estado em que é necessário uma quantidade maior de insulina do que o normal para obter uma resposta quantitativamente normal (YALOW;BERSON, 1960a, 1960b).

Em 1978, Ronald Kahn delineou a resistência à insulina como uma condição que ocorre sempre que concentrações normais de insulina resultam em uma resposta biológica abaixo do padrão normal (KAHN, 1978). Em 2010, um painel de especialistas debateu a maneira mais apropriada de avaliar a RI na infância e adolescência, que consistiu na caracterização como a redução da resposta dos tecidos às ações celulares mediadas pela insulina, representando o oposto da sensibilidade à insulina (LEVY-MARCHAL et al., 2010).

Na fase adulta a resistência à insulina é amplamente reconhecida como um fator crucial no desenvolvimento do diabetes mellitus tipo 2 (DM2), associando-se à obesidade, síndrome metabólica, hipertensão e doença cardíaca (REAVEN, 1997).

A DM2 e doenças cardiovasculares (DCV) emergem como uma preocupante epidemia global, demonstrando um grande aumento nas últimas décadas. Estima-se que o número total de pessoas com diabetes globalmente aumentará de 151 milhões em 2000 para 221 milhões até 2010 (AMOS et al., 1997) e para 300 milhões até 2025 (ZIMMET et al., 2001). Além disso, estudos mostram que tem ocorrido um aumento na identificação de DM2 e DCV na infância e adolescência (DE QUADROS et al., 2019; EKTA et al., 2016), sendo a RI o denominador comum entre DM2 e DCV (HAFFNER et al., 1992; BECK-NIELSEN et al., 1999).

2.2 Sensibilidade e Resistência Insulínica na Puberdade e Adolescência

Estudos recentes demonstram que há uma tendência de início precoce da

puberdade e menarca nas meninas, tendência essa que pode ser explicada pelo aumento da prevalência da obesidade visto que o IMC mais alto leva ao início mais precoce da puberdade. Essa tendência não foi observada em meninos (HERMAN et al., 1997).

Crianças pré-pubescentes crescem em média 5-6cm/ano e durante a puberdade essa velocidade é aumentada com uma rápida taxa de crescimento. Esse crescimento linear ocorre em resposta ao hormônio do crescimento. Os esteroides sexuais também contribuem para o aumento do hormônio do crescimento, promovendo esse crescimento ósseo longitudinal em ambos os sexos (TANNER et al., 1976; FRISCH et al., 1971).

Durante a puberdade, verifica-se uma diminuição da sensibilidade à insulina. Esta diminuição da sensibilidade à insulina foi também confirmada por um estudo semilongitudinal, que demonstra que esta diminuição ocorre durante a puberdade (WEITS et al., 1988). Durante a puberdade, é necessária uma concentração elevada de insulina para promover um crescimento ideal, o que resulta numa resistência insulínica fisiológica, um processo que é autolimitado, ou seja, uma condição temporária que ocorre naturalmente nos adolescentes (ZEITLER; NADEAU, 2020).

Os marcadores cardiovasculares ligados à sensibilidade à insulina se mostraram afetados pela puberdade. A pressão arterial sistólica aumenta em ambos os sexos, mas a uma taxa mais acelerada em meninos. A pressão arterial diastólica também aumentou em ambos os sexos, mas sem diferença entre eles (SCHWARTZ et al., 2008).

2.3 Obesidade na Puberdade e Adolescência

A obesidade afeta populações independentemente do estágio do ciclo da vida ou da condição socioeconômica. Deve ser considerada uma doença crônico-degenerativa e multifatorial, que apresenta fatores etiológicos reversíveis e irreversíveis e que pode ser apontada como causa direta de doenças graves (WAJCHENBERG et al., 2014)

Além disso, a obesidade é um dos contribuintes mais conhecidos para a resistência insulínica, mostra-se como um problema prevalente em diversos países. Em crianças e adolescentes, a obesidade é avaliada através do IMC, que é a medida

de triagem aceita, e tem sido utilizado para estudos na atualidade. Considera-se o cálculo do IMC/idade pelo z score com as curvas percentílicas acima de $z + 1$ (percentil 85) classificadas como sobrepeso e de $z + 2$ (percentil 95) como obesidade (REINEHR et al., 2015; MUST et al., 2005).

Nos Estados Unidos, aproximadamente 32% das crianças de 2 a 19 anos são consideradas com sobrepeso ou obesidade, e 17% são consideradas obesas. Isso aponta que cerca de um terço estão em risco de complicações associadas à resistência insulínica (OGDEN et al., 2014). A combinação de obesidade e puberdade representam grande grau de risco para os adolescentes no desenvolvimento da RI fisiológica, pois é caracterizada pela diminuição da oxidação da glicose e pela redução da supressão mediada pela insulina na oxidação de ácidos graxos livres. Com isso surge a necessidade aumentada de oxidação de gordura em decorrência da obesidade. Esse aumento da oxidação de gordura pode amplificar as manifestações glicêmicas da RI (HANNON et al., 2006; WU et al., 2006).

Na puberdade nota-se que a sensibilidade à insulina é inversamente proporcional ao índice de massa corporal (IMC) e percentual de gordura corporal, ou seja, os adolescentes obesos têm menor sensibilidade à insulina quando comparados com adolescentes de peso normal (LEVY-MARCHAL et al., 2010). Além disso, na adolescência a ação da insulina no metabolismo da glicose é cerca de 30% inferior, compensada pelo aumento da secreção de insulina (ARSLANIAN et al., 1994; CAPRIO et al., 1989).

2.4 Aspectos de Estilo de Vida: Tabagismo, Álcool, Atividade Física e Comportamento Sedentário

O uso de cigarros está associado ao aumento no triglicerídeo plasmático e diminuições na concentração plasmática de colesterol HDL. Tais mudanças são secundárias à resistência e captação de glicose, estimuladas por insulina e hiperinsulinemia (FACCHINI et al., 1992). No Brasil, segundo o estudo realizado pela Pesquisa Nacional de Saúde Escolar 2019 (PeNSE 2019), o consumo de cigarros entre adolescentes, de 13 a 17 anos, apresentou uma prevalência de 5,4%, quando os adolescentes foram perguntados se fumaram algum dia nos 30 dias anteriores à pesquisa (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2019).

Segundo estudo realizado por Choi (2019), indivíduos que aumentaram o consumo de bebida alcoólica tiveram um alto risco de síndrome metabólica (aOR, 1,45; IC 95%, 1,09 a 1,92) quando comparados a indivíduos que diminuíram o consumo de bebida alcoólica (aOR, 0,61; IC 95%, 0,44 a 0,84). No Brasil, segundo o estudo PeNSE 2019, o consumo de bebida alcoólica em adolescentes que tomaram pelo menos um copo ou uma dose nos 30 dias anteriores à pesquisa, apresentou uma prevalência de 25,3% (IC de 23,5% a 27,1% para IC de 95%) para adolescentes entre 13 a 17 anos.

O grau de atividade física afeta a resistência insulínica, pois na prática de atividade física a utilização de glicose em todo o corpo é aumentada, e essa captação é mediada pela insulina. Até mesmo uma atividade física moderada aumenta a captação de glicose pelas células musculares por mecanismos independentes da insulina (BURSTEIN et al., 1985).

O exercício modula a translocação dos transportadores GLUT4, que normalmente são controlados pela secreção de insulina (WEITS et al., 1988), aumentando a captação de glicose de forma aditiva aos mecanismos mediados por insulina, contribuindo para uma homeostase glicêmica mais eficiente no geral (AMIEL et al., 1986). A prática de atividade física melhora a sensibilidade à insulina. Um estilo de vida que incorpore exercícios aeróbicos e/ou atividade física que cumpram as diretrizes de intensidade moderada por pelo menos 30 minutos em 3-5 dias por semana está associado à melhora da sensibilidade à insulina e ao controle glicêmico. Além disso, os benefícios da atividade física e exercício são evidentes em todas as idades, desde crianças até adultos mais velhos, incluindo aqueles categorizados como assintomáticos/saudáveis, pré-diabéticos/síndrome metabólica, e pacientes com DM2 (BIRD et al., 2017).

O comportamento sedentário é caracterizado pela escassez no gasto de energia na realização de atividades sentado, reclinado ou em repouso (WORLD HEALTH ORGANIZATION, 2020b). Adultos com DM2 que substituíram várias horas por dia de sedentarismo por ficar em pé ou caminhar levemente apresentaram uma redução significativa da glicemia ao longo de 24 horas, em comparação com adultos que não substituíram os períodos de sedentarismo por atividade. Ao longo de 4 a 5 dias, a sensibilidade à insulina também melhorou nos grupos que substituíram os períodos

de sedentarismo por exercício. Essas mudanças parecem ocorrer por mediação da captação de glicose induzida pela contração, ao invés da captação de glicose mediada pela insulina. Assim, a melhora na sensibilidade à insulina pode ser observada até mesmo em um curto período de aumento dedicado de atividade não extenuante (DUVIVIER et al., 2016; DUNSTAN et al., 2012). De acordo com dados da PeNSE de 2019, a prevalência de tempo de tela sedentário em adolescentes de 13 a 17 anos que ficaram mais de 3 horas por dia nos 7 semanas anteriores à pesquisa em atividades sentado, de frente para telas de televisão, computador, celular foi de 61,6% (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2019).

2.5 Métodos de detecção de Resistência Insulínica

Nas últimas décadas, estudos de investigação metabólica clínica são dedicados ao desenvolvimento de métodos para medir a resistência à insulina e sua utilização em mecanismos teciduais e celulares responsáveis subjacentes (SIMONSON, 2014). Mais de quarenta anos se passaram desde a publicação de dois métodos importantes para estudar resistência à insulina: clamp euglicêmico hiperinsulinêmico e clamp hiperglicêmico (DEFRONZO et al., 1979).

Apesar do tempo, essas técnicas ainda são consideradas métodos padrão-ouro, porque efetivamente interrompem os mecanismos normais de retroalimentação homeostático que existem entre captação de glicose no tecido, secreção de insulina e a concentração plasmática de glicose, permitindo que o pesquisador estude cada variável de forma independente, enquanto controla as demais (DEFRONZO et al., 1979; SIMONSON, 2014). Contudo, a estimativa direta da resistência insulínica por meio dessas técnicas é complicada, trabalhosa, cara e raramente utilizada em grandes estudos epidemiológicos (SINGH et al., 2010).

Também existem as medidas indiretas de resistência à insulina. Elas podem ser divididas em três categorias, sendo a primeira derivada de amostra de jejum (HOMA-IR, QUICKI, FIRI, razão de insulina de jejum por glicose e razão de glicose de jejum por insulina); a segunda é derivada de valores de teste de tolerância oral a glicose (STUMVOLL et al., 2000) e a terceira é baseada em marcadores clínicos, medidas antropométricas e exames laboratoriais de jejum, como IMC, circunferência da cintura, diâmetro abdominal sagital, índice de triglicerídeos/glicose (TyG) e Razão de

triglicérides/HDL (Tg/HDL) (ANTUNA-PUENTE et al., 2011; IWANI et al., 40 2019; KIM, 2018; RUDVIK et al., 2018). Essas medidas se mostram mais aplicáveis a estudos de larga escala devido a sua simplicidade de aplicação (ARSLANIAN, 2005; MICHALISZYN et al., 2020).

Por fim, também existe o método HOMA-IR, que é um modelo matemático de interação de glicose e insulina utilizado para indicar o grau de hiperglicemia e, desse modo, indicar o grau de resistência insulínica e de deficiência das células beta-pancreáticas de acordo com as concentrações plasmáticas em jejum de glicose e de insulina do paciente (MATTHEWS et al., 1985).

2.6 Modelos de Aprendizado de Máquina

2.6.1 Modelagem preditiva de dados clínicos

No cotidiano a equipe médica e os responsáveis por políticas de saúde precisam fazer previsões sobre o prognóstico de uma doença. Também compete a equipe médica a análise da probabilidade de uma doença subjacente, em outras oportunidades é necessário avaliar as previsões para solicitações de testes e/ou terapias para grupo de pacientes em alto risco. Na área da saúde chamamos este processo de tomada de decisão clínica. Tradicionalmente, os profissionais da saúde utilizam a ferramenta da medicina baseada em evidências visando evitar adotar práticas de saúde que não sejam eficazes para o paciente. Além disso, a medicina baseada em evidências visa evitar tratamentos sem base científica comprovada (OXFORD Centre for Evidence-Based Medicine, 2024).

Nas últimas décadas, o estilo tradicional de tomada de decisão médica em que os médicos assumem a responsabilidade exclusiva pelas decisões de tratamento tem sido desafiado e nota-se a presença de tentativas para promover a tomada de decisão compartilhada, onde os pacientes têm a oportunidade de expressar seus valores e preferências e de participar das decisões sobre seus cuidados (COULTER, 1997).

Diante dessa nova onda de transformação, os modelos de previsão clínica podem contribuir, fornecendo dados para a tomada de decisões partilhadas e munindo os profissionais de saúde de informações e predições, para que estes possam decidir o tratamento de cada paciente com base em mais dados (GERDS et al., 2022).

2.6.2 Desenvolvimento, avaliação e comparação de modelos preditivos

Embora a Inteligência artificial (IA) e o aprendizado de máquina estejam estreitamente ligados, eles não são termos idênticos. A IA é um campo amplo da ciência que visa criar sistemas capazes de realizar tarefas que normalmente requerem inteligência humana, como reconhecimento de fala, tradução de idiomas, tomada de decisões, bem como tarefas criativas como gerar imagens e textos. Por outro lado, o aprendizado de máquina é um subcampo da IA que se concentra no desenvolvimento de algoritmos que permitem aos computadores aprender a partir de dados e assim realizar previsões em casos novos ou não vistos durante o treinamento (BURKOV, 2019; SARKER, 2021; PANCH, 2018).

Na etapa de treinamento, o modelo é treinado com o conjunto de dados e faz previsões. Em seguida, o modelo ajusta seus parâmetros com base em uma métrica que quantifica quão boas ou ruins são as previsões (função de perda). Também é de praxe, que durante o treinamento o modelo seja aplicado a um grupo “invisível” de observações (conjunto de validação) para obtermos uma avaliação mais confiável do desempenho em dados não vistos (GERDS, 2022).

Antes da escolha do modelo a ser utilizado, há uma etapa anterior que consiste na escolha do tipo de aprendizagem a ser utilizado. Atualmente existem diversos métodos de aprendizagem, mas três são bem utilizados na área de medicina: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem semi supervisionada (SARKER, 2021).

A aprendizagem supervisionada é um processo que seleciona recursos de entrada (rótulos) e em seguida atribui pesos para vincular os dados de entrada em um determinado valor (modelos de regressão) ou classe (modelos de classificação). A aprendizagem não supervisionada, consiste na análise e agrupamento de dados não rotulados. Com essa análise os modelos identificam padrões ocultos nos dados e os agrupa em conjuntos (SARKER, 2021). Por fim, o aprendizado semi supervisionado é uma abordagem que surgiu buscando mesclar as outras aqui citadas, e essa técnica consiste em processar dados não rotulados e cria representações-chave nas suas previsões (JING, 2021).

Seja supervisionado, não supervisionado ou semi supervisionado, o processo de criação dos modelos de aprendizado de máquina requer um conjunto de regras e

técnicas estatísticas que podem aprender padrões significativos a partir de dados, conhecidos como algoritmos (OIKONOMOU et al., 2023).

2.6.3 Principais algoritmos de aprendizagem de máquina em predição clínica

2.6.3.1 Modelos lineares generalizados (MLG)

Os modelos lineares generalizados são uma classe de modelos estatísticos que generalizam a regressão linear tradicional para situações onde os dados não seguem uma distribuição normal e/ou a relação entre variáveis não é linear. Esse tipo de modelo permite a criação de diversos modelos como regressão logística e regressão de *Poisson* (MÜLLER, 2004).

$$\text{Regressão Poisson: } y = e^{\left(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n \right)}$$

$$\text{Regressão logística: } y = \frac{1}{e^{\left(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n \right)}}$$

2.6.3.2 Decision Tree

Os modelos de *Decision Tree* se assemelham a um fluxograma de tomada de decisão. Esse tipo de modelo permite a visualização da tomada de decisão pois tem uma representação visual das suas decisões e possíveis consequências. Esse tipo de modelo toma decisões repartindo a árvore em nós, onde cada nó representa uma característica e cada ramificação representa uma possível resposta para essa característica. Esses modelos podem ser para resolver problemas de classificação ou de regressão (QUINLAN, 1986).

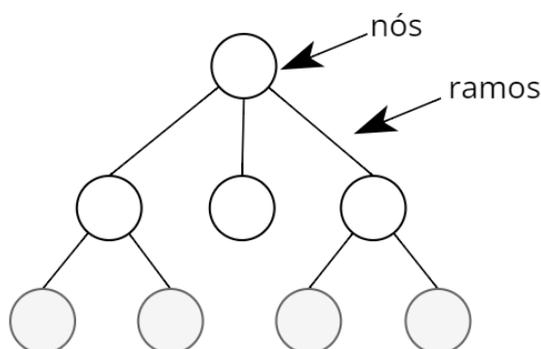


Figura 1. Desenho esquemático do modelo *decision tree* contendo 8 nós e 7 ramos. Fonte: Imagem criada pela autora.

2.6.3.3 Random Forest

Os modelos *Random Forest* são uma combinação de diversos modelos de árvore de decisão. Nesse tipo de modelo, não há uma árvore individual que seja escolhida como a melhor árvore para o resultado da previsão. Em vez disso, o resultado final da previsão é determinado pela combinação das previsões de todas as árvores presentes no algoritmo (BREIMAN et al., 2001).

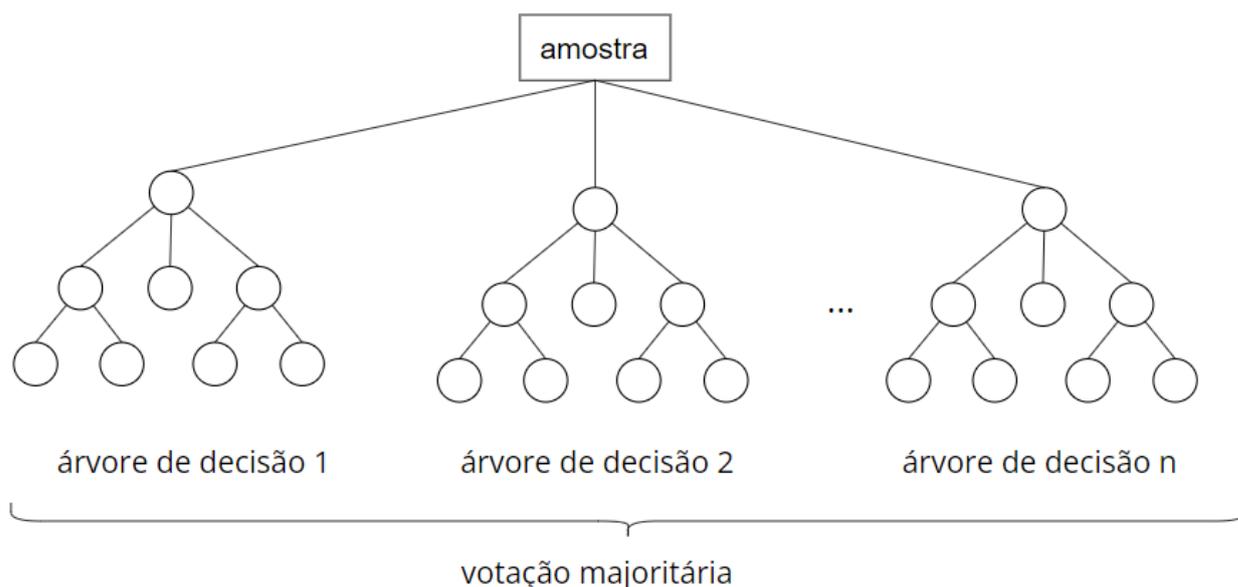


Figura 2. Desenho esquemático do modelo *random forest* demonstrando a presença de diversas árvores de decisão. Fonte: Imagem criada pela autora.

2.6.3.4 Support Vector Machine (SVM)

O modelo SVM é um algoritmo capaz de realizar tarefas de classificação e regressão. A SVM busca encontrar um hiperplano que melhor separa diferentes classes de dados. Esse hiperplano é escolhido de forma a maximizar a margem, que é a distância entre os pontos mais próximos de cada classe (EVGENIOU et al., 2001).

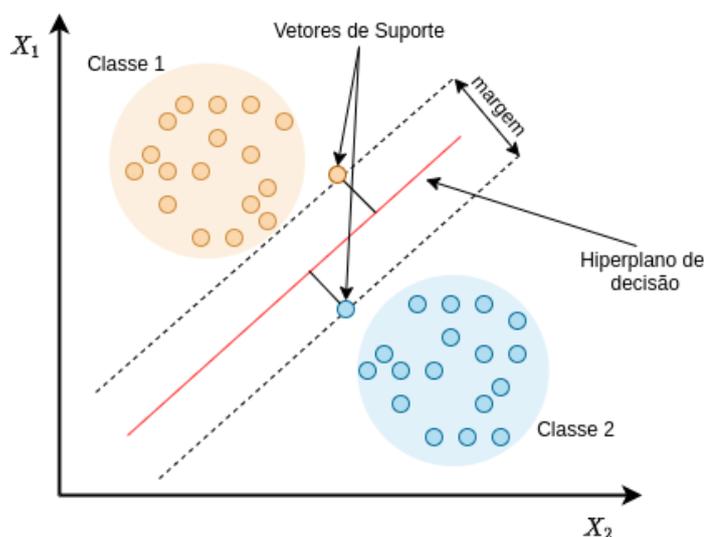


Figura 3. Estrutura geral do algoritmo SVM. O algoritmo define o hiperplano de separação entre classes, onde a seleção desses vetores de suporte visam maximizar a margem de separação entre as classes. Fonte: BARBOSA et al., 2021.

2.6.3.5 Redes Neurais Profundas

Redes neurais profundas, tradução de *Deep Neural Network* (DNN), são modelos de aprendizado de máquina compostos por múltiplas camadas, e em cada camada encontra-se presente a estrutura chamada de neurônio. Cada neurônio realiza uma operação matemática que recebe as entradas, realiza o processamento e gera saídas que são passadas para a camada a seguir. Cada camada aprende a representar os dados de forma progressivamente, começando com características mais simples e avançando para características mais complexas, conforme avançamos em sentido a camada de saída. Essa capacidade de aprender representações hierárquicas dos dados faz com que esse tipo de modelo seja eficaz em tarefas como reconhecimento de imagem, processamento de linguagem natural e previsão. Tais características tornam as redes neurais profundas uma excelente ferramenta para lidar com problemas complexos de aprendizado e problemas não lineares (GOODFELLOW et al., 2016).

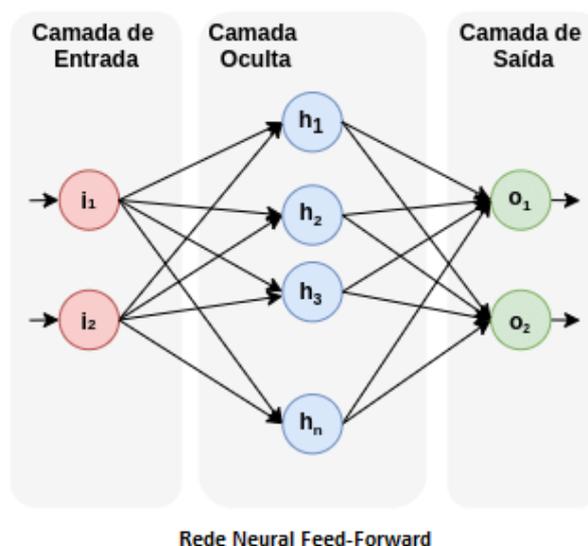


Figura 4. Camadas de uma Rede Neural e interação entre os neurônios. A rede neural mostrada é formada por três camadas, sendo elas a de entrada, oculta e saída. Fonte: BARBOSA et al., 2021.

2.6.3.6 Extreme gradient boosting (XGBoost)

O modelo XGBoost pertence à família de algoritmos denominada de *ensemble*. Esse conceito consiste na utilização de vários modelos mais simples (como por exemplo árvores de decisão) que são combinados para formar um modelo. O XGBoost utiliza o método de *Gradient Boosting* que busca adicionar modelos sequencialmente, onde cada novo modelo tenta corrigir os erros dos anteriores, através do ajuste dos pesos para minimizar a função de perda. O XGBoost também é reconhecido por sua eficiência computacional e por lidar com grandes volumes de dados, realizando o treinamento de forma rápida (CHEN et al., 2016)

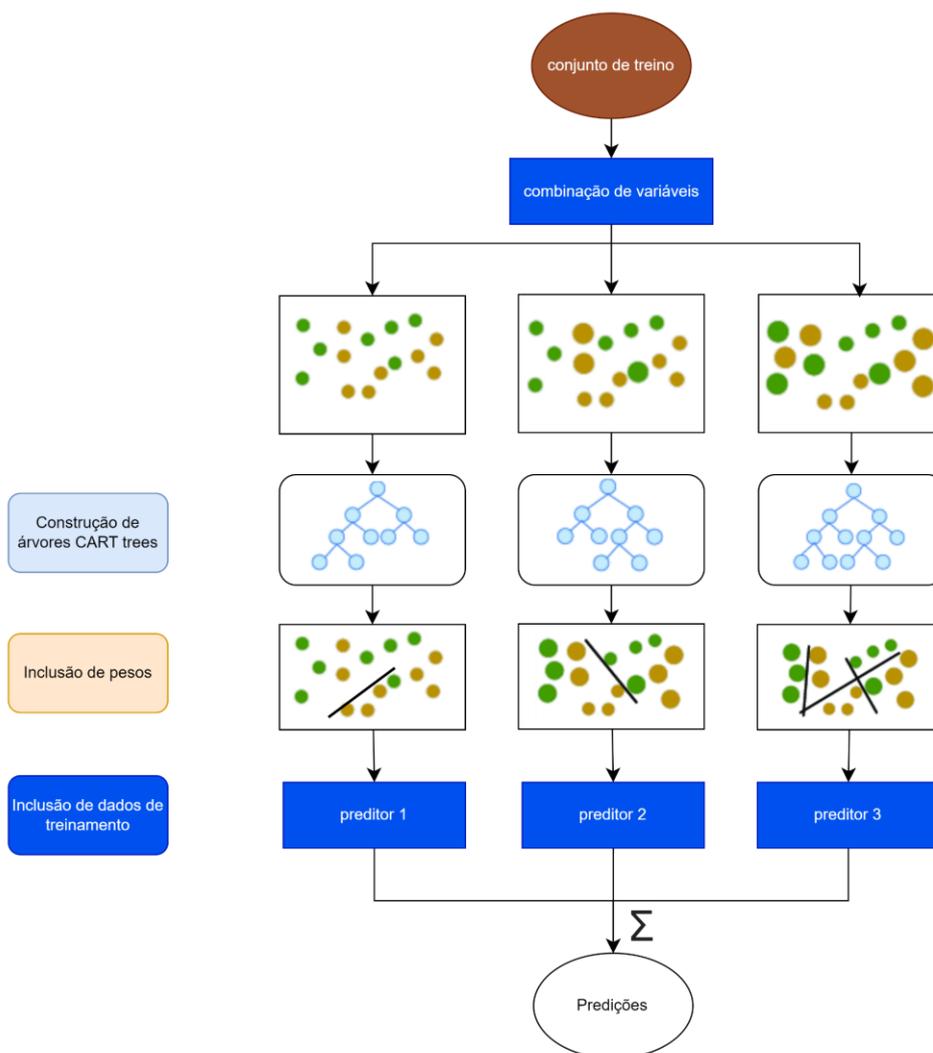


Figura 5. Exemplo de um modelo XGBoost contendo 3 árvores de decisão que buscam minimizar a função de perda para obtenção do melhor resultado. Fonte: Imagem criada pela autora.

2.6.3.7 Categorical Boosting (CatBoost)

O modelo CatBoost em termos de arquitetura se assemelha bastante do modelo XGBoost. A grande diferença é que o CatBoost pode consumir variáveis categóricas diretamente, sem a necessidade de um pré-processamento extensivo como a codificação one-hot que transforma variáveis categóricas em numéricas (PROKHORENKOVA et al.,2021).

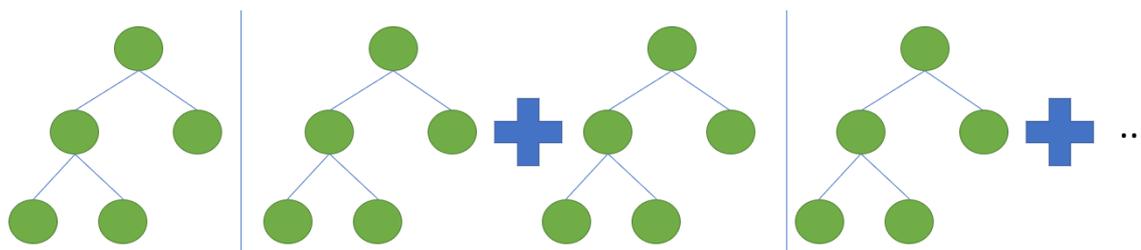


Figura 6. Exemplo de um modelo Catboost contendo 3 árvores que buscam minimizar a função de perda para obtenção do melhor resultado. Fonte: Imagem criada pela autora. Fonte: SHAHANI et al., 2022.

2.7 Modelos de predição de resistência insulínica

Na última década houve um aumento na criação e pesquisa de modelos de aprendizado de máquina buscando auxiliar no diagnóstico, tratamento personalizado e prognóstico de doenças. Em um estudo chinês com crianças de 6 a 12 anos, cinco modelos de aprendizado de máquina foram avaliados para prever RI (ZHANG et al., 2022): Regressão Logística, *Support Vector Machine*, *Random Forest*, XGBoost e Catboost. Dentre esses, o XGBoost demonstrou o melhor desempenho com uma AUC de 0,85.

De forma semelhante, outro estudo chinês envolvendo adultos com mais de 40 anos treinou sete modelos de aprendizado de máquina (ZHANG et al., 2024): RL, SVM, RF, *ExtraTrees*, *Light Gradient-Boosting Machine* (LightGBM), XGBoost e Árvore de Classificação e Regressão CART. O LightGBM exibiu desempenho superior com uma AUC de 0,86.

Um estudo coreano com adultos acima de 40 anos, sete modelos de aprendizado de máquina foram avaliados (PARK et al., 2022): RL, XGBoost, *Decision Tree*, *K-Nearest Neighbors* (KNN), SVM, RF e Rede Neural Profunda (RNP). Tanto o RL quanto o XGBoost apresentaram os melhores resultados, cada um com uma AUC de 0,86.

Um estudo americano com pacientes com doença renal crônica acima de 18 anos avaliou quatro modelos de aprendizado de máquina (LEE et al., 2022): XGBoost, Regressão Logística, Rede Neural Profunda (RNP) e RF. O XGBoost foi o mais eficaz com uma AUC de 0,78.

2.8 Preditores associados a resistência insulínica

Estudos recentes identificaram preditores associados a resistência insulínica em adolescentes e adultos jovens, incluindo obesidade (LEVY-MARCHAL et al., 2010), baixo colesterol HDL (HOWARD et al., 1998; REYES-SOFFER et al., 2015), alto colesterol LDL, alto triglicérides (SATTAR et al., 2015), elevada circunferência da cintura (RAMÍREZ-MANENT et al., 2023; DEUSDARÁ et al., 2022) e pressão arterial alta (MORAIS et al., 2016). Além disso, fatores de estilo de vida como comportamento sedentário (TIMMERMAN et al., 2018; YARIBEYGI et al., 2019), consumo de álcool (CHOI et al., 2019) e bebidas açucaradas (MA et al., 2016) estão ligados ao aumento da RI. Foi observado que fumantes exibem maior resistência à insulina (FACCHINI et al., 1992). Em relação aos fatores de características socioeconômicas, embora a renda familiar não tenha afetado diretamente o HOMA-IR, a escolaridade dos pais teve uma forte relação com a RI presente nos adolescentes do estudo (GOODMAN et al., 2007).

Por outro lado, o exercício físico melhora a sensibilidade à insulina (DELA et al., 1993; FROSIG et al., 2007). Uma revisão sistemática avaliou o desempenho metodológico de estudos que previam o risco de diabetes mellitus tipo 2 (DM2) não diagnosticado ou risco futuro. Nessa revisão, os preditores mais comuns foram idade, histórico familiar de diabetes, índice de massa corporal, hipertensão, circunferência da cintura e sexo (COLLINS et al., 2011).

3. JUSTIFICATIVA

Fatores de riscos associados à resistência insulínica em crianças e adolescentes incluem doenças cardiovasculares, diabetes mellitus tipo 2 e hipertensão (ZAVARONI et al.,1999). Em 2015, o diabetes em adultos custou cerca de 1,4 trilhões de dólares no mundo todo, representando quase 2% do PIB mundial, de acordo com Bommer et al. (2018). O estudo prevê que até 2030, esse custo aumentará para até 2,5 trilhões de dólares.

Portanto, diante desses aspectos, nota-se a importância de identificar a resistência insulínica em crianças e adolescentes visando melhorar a qualidade de vida, bem como do ponto de um melhor aporte dos investimentos do cuidado em saúde nessa população alvo (ZEITLER; NADEAU, 2020;LEVY-MARCHAL et al., 2010).

4. HIPÓTESE

Modelos de aprendizado de máquina superam modelos clássicos, como a regressão logística, na previsão da resistência à insulina. Isso ocorre porque eles conseguem lidar melhor com grandes volumes de dados e capturar padrões complexos e interações não lineares entre as variáveis.

5. OBJETIVO

5.1 Objetivo geral

Avaliar o modelo de aprendizado de máquina com melhor performance para prever a resistência insulínica em adolescentes do Estudo de Riscos Cardiovasculares em Adolescentes.

5.2 Objetivos específicos

Analisar modelos de *machine learning* utilizados para predição em resistência insulínica

Desenvolver um modelo de inteligência artificial que poderá ser usado no futuro para identificar a resistência insulínica em adolescentes.

Analisar o modelo de melhor performance com base na melhor métrica de acurácia.

6. METODOLOGIA

6.1 Estudo de Riscos Cardiovasculares em Adolescentes

6.1.1 Desenho amostral

O ERICA é um estudo multicêntrico nacional, realizado entre fevereiro de 2013 e novembro de 2014, cujo objetivo foi fornecer estimativas de fatores de risco cardiovascular, de síndrome metabólica e seus componentes em adolescentes (DA SILVA et al., 2016; BLOCH et al., 2015).

No estudo foram avaliadas as condições de saúde de cerca de 75 mil estudantes entre 12 e 17 anos, de 1.247 escolas brasileiras, públicas e particulares, distribuídas por 122 cidades participantes. Para a seleção amostral dos adolescentes foi utilizado o Censo Escolar de 2011 (VASCONCELLOS et al., 2015).

A população utilizada foi estratificada em 32 estratos geográficos, sendo 27 capitais e cinco estratos com municípios com mais de 100.000 habitantes de cada uma das cinco macrorregiões. O processo de amostragem foi realizado em três estágios, sendo eles escola, turma e alunos. Em cada escola, foram escolhidas três combinações de turno (manhã x tarde) e ano, correspondentes aos três últimos anos do ensino fundamental e os três anos do ensino médio (VASCONCELLOS et al., 2015).

No total, foram selecionadas 1.251 escolas de 124 municípios do total de 273 municípios elegíveis. No estudo houve perdas de 4 escolas, sendo 3 no Estado de São Paulo e 1 no Estado do Amapá, resultando na coleta de dados em 1.247 escolas de 122 municípios. Os resultados têm representatividade nacional para os 32 estratos e as 5 macrorregiões (BLOCH et al., 2015; VASCONCELLOS et al., 2015). Uma descrição detalhada do projeto do estudo foi publicada, bem como todas as informações do questionário do ERICA (BLOCH et al., 2015).

Apesar do estudo analisar as condições de saúde de cerca de 75 mil estudantes, na amostra utilizaremos somente as variáveis presentes no subconjunto de 37.815 adolescentes, pois havia uma necessidade de jejum de 12 horas para a coleta dos exames bioquímicos. Com isso o perfil dos adolescentes que correspondeu a essa condição foram os alunos do turno da manhã (DA SILVA et al., 2016).

6.1.2 Questionários

O questionário do adolescente foi autopreenchido pelos estudantes em coletor eletrônico de informações, *Personal Digital Assistant*, modelo LG GM750Q. O questionário possuía 11 blocos temáticos (características sociodemográficas, trabalho e emprego, atividade física, comportamento alimentar, tabagismo, consumo de álcool, saúde reprodutiva, saúde bucal, morbidade referida, duração do sono e transtorno mental comum) contendo um total de 105 questões (DA SILVA et al., 2016).

O questionário do responsável possuía questões sobre problemas de saúde na família, além de outras variáveis relacionadas ao adolescente. Já o questionário da escola consistia em analisar características físicas encontradas em cada escola, como cantina, bebedouros e quadras de esportes. Todos os alunos participantes do estudo e seus pais ou responsáveis legais assinaram um termo de consentimento (BLOCH et al., 2015).

6.1.3 Recordatório e exames bioquímicos

O recordatório consistiu em uma entrevista sobre a alimentação do adolescente nas últimas 24 horas. Para os dados provenientes dos exames bioquímicos foi realizada uma coleta de sangue em uma subamostra contendo cerca de 40 mil adolescentes, pois era necessário o jejum de 12 horas. Logo, somente os alunos do turno da manhã participaram dos exames bioquímicos (DA SILVA et al., 2016).

6.1.4 Medidas antropométricas e pressão arterial

As medidas antropométricas coletadas foram o peso, estatura e perímetro da cintura. Para os dados da pressão arterial, realizou-se duas medidas utilizando o aparelho, marca Omron®, modelo 705-CP, validado para uso em adolescentes (BLOCH et al., 2015; STERGIOU et al., 2006)

6.2 Conjunto de dados e seleção de variáveis

6.2.1 Colesterol HDL

A variável colesterol HDL representa a medida do nível plasmático de HDL (mg/dL) em jejum (BLOCH et al., 2015).

6.2.2 Colesterol LDL

A variável colesterol LDL representa o nível plasmático de LDL em jejum (mg/dL). O LDL foi calculado pela equação de Friedewald [colesterol_hdl + (triglicerídeos/5)], apenas para aqueles com colesterol menor que 400 mg/dl (FRIEDEWALD et al., 1972).

6.2.3 Triglicerídeos

A variável triglicerídeos representa o nível plasmático de triglicerídeos em jejum (mg/dL) (ALBERTI; ZIMMET, 2007).

6.2.4 Circunferência da cintura

A variável circunferência da cintura foi medida duas vezes no ponto médio entre a curva inferior da última costela fixa e a curva superior da crista íliaca, com o adolescente em pé, braços ao lado do corpo, pés juntos e abdômen relaxado. Foi permitida uma variação máxima de um centímetro entre as duas medições. O assistente pessoal digital calculava então automaticamente a média das duas medidas (BLOCH et al., 2015).

6.2.5 Inatividade Física

A variável inatividade física foi criada a partir da variável "tempo despendido em desporto em min/semana", que inclui atividades desportivas como futebol, futsal, handebol, basquetebol, patins, atletismo, natação, ginástica olímpica, luta, dança, corrida, ciclismo, caminhada, voleibol de quadra, voleibol de praia, queimada, surf, musculação, ginástica, ténis de quadra, ginástica e futebol de praia. O adolescente foi considerado inativo se o tempo gasto na atividade foi menor ou igual a 420 min/semana (WORLD HEALTH ORGANIZATION, 2010).

6.2.6 Hábitos de fumo

A variável hábitos de fumo foi criada a partir da variável "nos últimos 30 dias, em quantos dias você fumou cigarros?" Um adolescente foi classificado como fumador se tivesse fumado pelo menos um cigarro nos últimos 30 dias (WARREN et al., 2008).

6.2.7 Consumo de álcool

A variável consumo de álcool foi criada a partir da variável "bebedor", que classifica se uma pessoa consome álcool ou não (consumiu pelo menos 1 copo de álcool nos últimos 30 dias). Foram classificados como consumidores de álcool os adolescentes que responderam ter consumido pelo menos um copo de bebida alcoólica nos últimos 30 dias, excluindo-se aqueles que não consumiram ou não se lembravam (WARREN et al., 2008).

6.2.8 Comportamento sedentário

A variável tempo de tela foi criada para mapear o comportamento sedentário dos adolescentes. O bloco sobre incluiu perguntas sobre as horas de uso de computador, televisão e videogame em um dia médio da semana. Foram considerados sedentários os adolescentes que responderam que utilizam esses aparelhos por mais de 3 horas por dia (AMERICAN ACADEMY OF PEDIATRICS, 2001).

6.2.9 Variável socioeconômica

A variável tipo de escola foi criada para representar os fatores socioeconômicos dos adolescentes, sendo que a variável possuía os níveis de escola pública e privada (GONZAGA et al., 2012).

6.2.10 Estado nutricional

A variável estado nutricional foi criada a partir do IMC, definido como o peso em quilogramas dividido pelo quadrado da altura em metros. A classificação do estado nutricional foi determinada pela Organização Mundial de Saúde, que leva em consideração o IMC por idade e sexo para adolescentes, utilizando escores Z. Os estados nutricionais de obesidade, sobrepeso, normal, abaixo do peso e baixo peso, foram definidos da seguinte forma: Z-score $> +2$; Z-score $> +1$ e Z-score $\leq +2$; Z-score ≥ -2 e Z-score $\leq +1$; Z-score < -3 ; Z-score ≥ -3 e Z-score < -2 , respectivamente (DE ONIS et al., 2007).

6.2.11 Pressão Arterial

A variável pressão arterial (PA) foi criada com base nos critérios de classificação do *Fourth Report on the Diagnosis, Evaluation, and Treatment of High Blood Pressure in Children and Adolescents* (NATIONAL HIGH BLOOD PRESSURE EDUCATION

PROGRAM WORKING GROUP ON HIGH BLOOD PRESSURE IN CHILDREN AND ADOLESCENTS, 2004). Os níveis criados para a PA foram "Normal[=0]": Percentil da PA <P90, "Limítrofe[=1] (pré-hipertensão)": Percentil da PA P90 a <P95 ou PA≥120/80mmHg, "PA alta[=2] (hipertensão estágio 1)": Percentil da PA P95 a P99 + 5mmHg, "PA muito alta [=3] (hipertensão estágio 2)": percentil da PA >P99 + 5mmHg, "sem classificação da PA [=9]": Apesar de terem medições válidas da tensão arterial sistólica e da tensão arterial diastólica, não tinham uma medição da altura para verificar o percentil.

6.2.12 Modelo Homeostático de Avaliação da Resistência à Insulina

Chissini et al. (2020) analisaram o ERICA e mostraram que o ponto de corte ideal para o HOMA-IR associado à síndrome metabólica era de 2,8 para a população total, 2,32 para adolescentes do sexo masculino e 2,87 para adolescentes do sexo feminino. Com base nesse ponto de corte, foi criada uma variável binária para identificar a resistência à insulina.

6.3 Desenho do estudo

O conjunto alvo do estudo foi de 37.815 adolescentes, devido a necessidade de um período de jejum de 12 horas para a coleta de exames bioquímicos, condição cumprida apenas pelos alunos do turno da manhã. Neste estudo, estratificamos os dados por sexo e todas as variáveis foram coletadas por meio de questionários dos adolescentes, questionários dos pais, recordatório de 24 horas, exames bioquímicos, medidas antropométricas e pressão arterial (DA SILVA et al., 2016). As variáveis com até 1% de dados faltantes foram excluídas do conjunto de dados, resultando em um conjunto de dados de 37.454 adolescentes (LITTLE; RUBIN, 2019; EMMANUEL et al., 2021; REN et al., 2023).

Na análise descritiva de todo o conjunto de dados, foram calculadas frequências absolutas e relativas para as variáveis categóricas, medidas de tendência central (média) e de variação (desvio-padrão) para as variáveis contínuas com distribuição normal. Para as variáveis não contínuas foram calculadas medianas e intervalos interquartílicos (IQR).

Dividimos os 37.454 adolescentes em dois grupos, estratificados pelo sexo. Em seguida, cada subconjunto estratificado foi dividido aleatoriamente para formar os

conjuntos de dados de treino e de teste para os algoritmos, mantendo a proporção de 70% dos dados para treino e 30% para teste. Para garantir a generalização e evitar o sobreajuste, cada grupo (meninas e meninos) foi dividido para manter a mesma proporção da variável de resultado nos subconjuntos de treinamento e teste. Além disso, padronizamos as variáveis com escalas diferentes para garantir a uniformidade na escala de todas as variáveis preditoras, garantindo assim uma melhor convergência dos modelos desenvolvidos.

O treino dos modelos de aprendizagem de máquina utilizou um método de validação cruzada k-fold, dividindo o conjunto de dados de treino em dez estratos de independentes (KOHAVI, 1995). Um estrato foi utilizado como conjunto de validação e os outros nove para treino. Este processo foi repetido dez vezes, utilizando em cada rodada um estrato diferente como conjunto de validação. Todos os algoritmos foram avaliados no subconjunto reservado para teste, com o objetivo de produzir um algoritmo que possa generalizar e produzir bons resultados para uma amostra que não esteja presente na base de dados. O desempenho preditivo de cada modelo de aprendizado de máquina gerado foi avaliado utilizando a área sob a curva (AUC). Também analisamos outras métricas como a sensibilidade, especificidade, pontuação f1, valor preditivo positivo (VPP) e valor preditivo negativo (VPN).

Para avaliar a contribuição de cada variável para o modelo com melhor desempenho, utilizamos a abordagem *SHapley Additive exPlanations* (SHAP), que fornece informações sobre a importância preditiva de cada variável e o seu impacto (positivo ou negativo) na predição realizada (LUNDBERG et al., 2017). Também analisamos a curva de calibração das probabilidades previstas por cada modelo, buscando verificar se as previsões de probabilidade estavam bem calibradas e próximas da curva de calibração perfeita (OIKONOMOU et al., 2023). Finalmente, foi analisado o desempenho do melhor modelo em comparação com as estratégias de tratamento convencionais utilizando a análise da curva de decisão (DCA), tendo em conta a utilização do indicador clínico HOMA-IR para detetar a resistência à insulina (VICKERS et al., 2016). No caso da predição de resistência insulínica, a estratégia clínica poderia ser a modificação do estilo de vida, incluindo aconselhamento dietético, atividade física e perda de peso, bem como apoio profissional para ajudar nesta mudança.

Em seguida, realizamos uma análise de sensibilidade comparando a utilização do desenho amostral complexo e do desenho amostral não complexo no desenvolvimento do modelo de regressão logística. Realizamos uma análise de sensibilidade comparando o uso da amostragem complexa e não complexa no desenvolvimento do modelo de regressão logística. Para isso usamos o valor de corte HOMA-IR de 2,63 para meninas e 2,28 para meninos, que corresponde ao valor de corte do percentil 75.

Todas as análises descritivas e o desenvolvimento do algoritmo foram efectuados utilizando a linguagem de programação R com versão 4.1.2 (R Core Team, Vienna, Austria). O objetivo foi utilizar ferramentas de código aberto para minimizar os custos de implementação de projectos tecnológicos (THE R PROJECT FOR STATISTICAL COMPUTING, 2022).

6.4 Desenvolvimento e avaliação dos modelos de aprendizado de máquina

No presente estudo, foram desenvolvidos modelos de classificação binária para prever a resistência insulínica. Os algoritmos de classificação binária mais comumente usados foram testados: Regressão Logística, *Poisson*, *Decision Tree*, *Random Forest*, *Support Vector Machine*, *eXtreme Gradient Boosting* e Redes Neurais Profundas.

O conjunto de dados de 37.454 adolescentes foi estratificado por sexo, onde cada subconjunto foi dividido em um conjunto de treinamento (70%) e um conjunto de teste (30%), sendo que na divisão de treino e teste ocorreu a estratificação pela variável desfecho se manter proporcional entre ambos os conjuntos de treino e teste. Na etapa de pré-processamento as variáveis numéricas foram padronizadas para garantir que tivessem escalas comparáveis e que os modelos convergissem mais eficientemente para a resposta.

O conjunto de treinamento foi utilizado para definir os parâmetros e hiperparâmetros de cada modelo desenvolvido durante a fase de treinamento. O conjunto de teste foi usado para avaliar o desempenho do modelo treinado com dados não vistos anteriormente. Variáveis ausentes com quase 1% de valores foram removidas do conjunto de dados. Na etapa de treinamento, utilizamos uma validação cruzada k-fold com dez estratos de dados independentes para treinar cada modelo. A

Figura 1 mostra todas as etapas envolvidas na construção dos modelos de aprendizado de máquina.

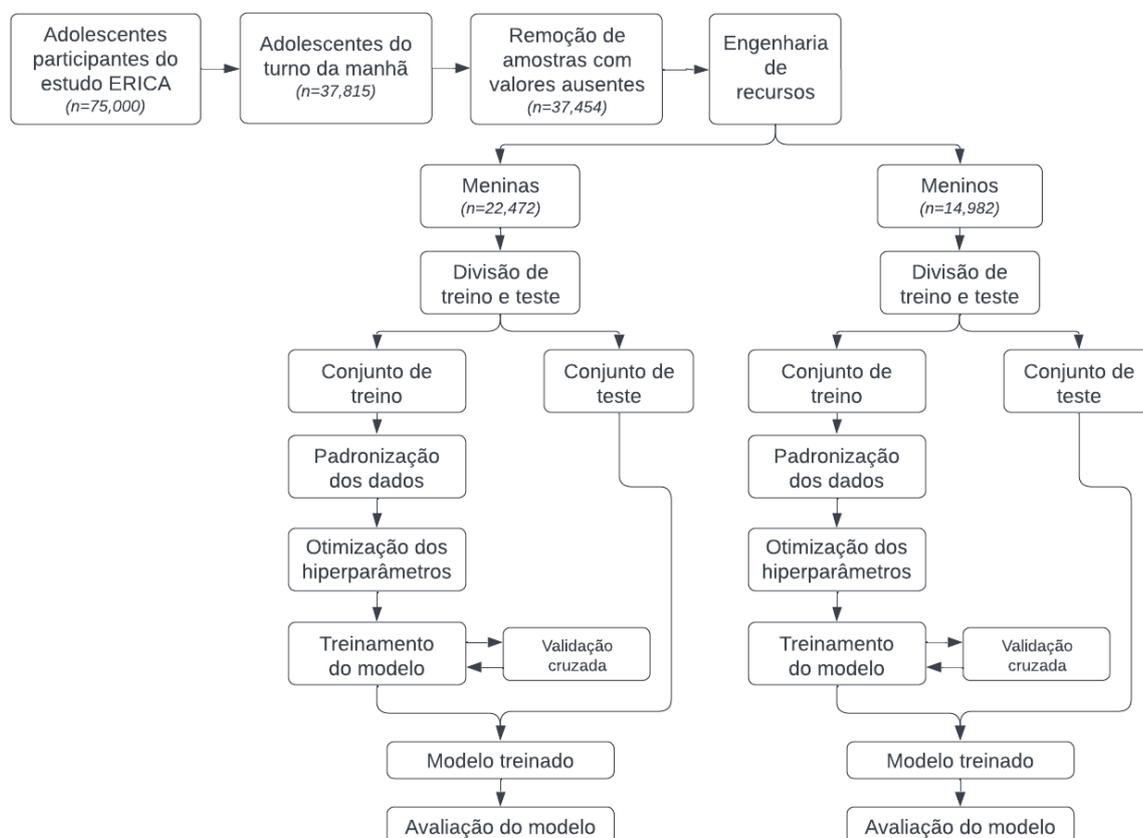


Figura 7. Etapas do fluxo de Aprendizado de Máquina desenvolvido para treinar modelos de aprendizado de máquina com o objetivo de prever a resistência à insulina no Estudo de Riscos Cardiovasculares em Adolescentes (ERICA) 2013-2014. . Dos 37.815 participantes potencialmente elegíveis para o desenvolvimento do modelo, 37.454 adolescentes atenderam aos critérios de inclusão.

6.5 Aspectos éticos

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Instituto de Estudos em Saúde Pública, Universidade Federal do Rio de Janeiro (Aprovação Número 45/2008) e pelos Comitês de Ética de cada participante instituição. Os comitês de ética locais aprovaram o estudo, e o consentimento informado foi obtido de todos os indivíduos e seus pais ou responsáveis legais incluídos neste estudo.

7. RESULTADOS

7.1 Manuscrito

8. CONSIDERAÇÃO FINAIS

REFERÊNCIAS

- ALBERTI, S. G.; ZIMMET, P. The IDF Consensus definition of the metabolic syndrome in children and adolescents. [S. l.]: **International Diabetes Federation**, 2007. 24 p.
- AMIEL, S. A.; SHERWIN, R. S.; SIMONSON, D. C.; LAURITANO, A. A.; TAMBORLANE, W. V. Impaired insulin action in puberty: a contributing factor to poor glycemic control in adolescents with diabetes. *New England Journal of Medicine*, v. 315, n. 4, p. 215–219, 1986.
- AMERICAN ACADEMY OF PEDIATRICS. Children, adolescents, and television. **Pediatrics**, v. 107, n. 2, p. 423-426, 2001.
- AMOS, A. F.; MCCARTY, D. J.; ZIMMET, P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. **Diabetologia**, v. 14, Suppl 5, p. S1–S85, 1997.
- Antuna-Puente, B.; Disse, E.; Rabasa-Lhoret, R.; Laville, M.; Capeau, J.; Bastard, J.P. How can we measure insulin sensitivity/resistance?. **Diabetes Metab.** 2011, 37, 179–188. <https://doi.org/10.1016/j.diabet.2011.01.002>
- ARSLANIAN, S. A.; KALHAN, S. C. Correlations between fatty acid and glucose metabolism: potential explanation of insulin resistance of puberty. **Diabetes**, v. 43, n. 7, p. 908-914, 1994.
- ARSLANIAN, S. A. Clamp techniques in paediatrics: what have we learned?. **Hormone Research**, v. 64, suppl. 3, p. 16-24, 2005. DOI: 10.1159/000089313.
- BARBOSA, G.; BEZERRA, G. M.; MEDEIROS, D. S. V. de; ANDREONI, M. Segurança em redes 5G: oportunidades e desafios em detecção de anomalias e previsão de tráfego baseadas em aprendizado de máquina. In: **MINICURSOS DO XXI SIMPÓSIO BRASILEIRO DE SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS**, 2021, São Paulo. Anais... São Paulo: Sociedade Brasileira de Computação, 2021. p. 145-189. DOI: 10.5753/sbc.7165.8.4.
- BECK-NIELSEN, H.; EUROPEAN GROUP FOR THE STUDY OF INSULIN RESISTANCE (EGIR). General characteristics of the insulin resistance syndrome: prevalence and heritability. **Drugs**, v. 58, p. 7-10, 1999.

BIRD, S. R.; HAWLEY, J. A. Atualização sobre os efeitos da atividade física na sensibilidade à insulina em humanos. **BMJ Open Sport & Exercise Medicine**, v. 2, n. 1, e000143, mar. 2017. DOI: 10.1136/bmjsem-2016-000143.

BLOCH, K. V. et al. The study of cardiovascular risk in adolescents – ERICA: rationale, design and sample characteristics of a national survey examining cardiovascular risk factor profile in Brazilian adolescents. **BMC Public Health**, v. 15, n. 1, p. 94, 7 dez. 2015.

BOMMER, C. et al. Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030. **Diabetes Care**, v. 41, n. 5, p. 963–970, maio 2018.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 01 jul. 2024.

BURSTEIN, R. et al. Reversão aguda da ação aumentada da insulina em atletas treinados. Associação com alterações do receptor de insulina. **Diabetes**, v. 34, n. 8, p. 756-760, ago. 1985. DOI: 10.2337/diab.34.8.756.

BURKOV, Andriy. The Hundred-Page Machine Learning Book. Edição ilustrada. Andriy Burkov, 2019. 141 p. ISBN 1999579518, 9781999579517.

CAMERON, A. J.; SHAW, J. E.; ZIMMET, P. Z. The metabolic syndrome: prevalence in worldwide populations. **Endocrinology and Metabolism Clinics of North America**, v. 33, p. 351-375, 2004.

CAPRIO, S. et al. Increased insulin secretion in puberty: a compensatory response to reductions in insulin sensitivity. **The Journal of Pediatrics**, v. 114, n. 6, p. 963-967, 1989.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. 2016. Disponível em: <https://doi.org/10.48550/arXiv.1603.02754>. Acesso em: 01 jul. 2024.

CHISSINI, R. DE B. C. et al. Cutoff values for HOMA-IR associated with metabolic syndrome in the Study of Cardiovascular Risk in Adolescents (ERICA Study). **Nutrition**, v. 71, 2020.

CHOI, S. et al. Association between change in alcohol consumption and metabolic syndrome: Analysis from the health examinees study. **Diabetes and Metabolism**

Journal, v. 43, n. 5, p. 615-626, 2019.

COLLINS, G. S. et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. **BMC Medicine**, v. 9, n. 1, p. 103, 8 dez. 2011.

COULTER, A. Partnerships with patients: the pros and cons of shared clinical decision-making. **Journal of Health Services Research & Policy**, v. 2, n. 2, p. 112-121, 1997.

DA SILVA, T. L. N. et al. Response rate in the study of cardiovascular risks in adolescents - ERICA. **Revista de Saúde Pública**, v. 50, n. suppl 1, p. 1s-13s, 2016.

DE ONIS, M. et al. Development of a WHO growth reference for school-aged children and adolescents. **Bulletin of the World Health Organization**, v. 43, n. 497, p. 660-667, 2007.

DE QUADROS, T. M. B. et al. Utility of anthropometric indicators to screen for clustered cardiometabolic risk factors in children and adolescents. **Journal of Pediatric Endocrinology and Metabolism**, v. 32, p. 49–55, 2019.

DEFRONZO, R. A.; TOBIN, J. D.; ANDRES, R. Glucose clamp technique: a method for quantifying insulin secretion and resistance. **American Journal of Physiology**, v. 237, n. 3, p. E214-23, set. 1979. DOI: 10.1152/ajpendo.1979.237.3. E214.

DELA, F. et al. Glut 4 and insulin receptor binding and kinase activity in trained human muscle. **The Journal of Physiology**, v. 469, n. 1, p. 615-624, 1 set. 1993.

DEUSDARÁ, R.; DE MOURA SOUZA, A.; SZKLO, M. Association between Obesity, Overweight, Elevated Waist Circumference, and Insulin Resistance Markers among Brazilian Adolescent Students. **Nutrients**, v. 14, n. 17, artigo 3487, ago. 2022. DOI: 10.3390/nu14173487.

DUNSTAN, D. W.; KINGWELL, B. A.; LARSEN, R.; HEALY, G. N.; CERIN, E.; HAMILTON, M. T.; et al. Breaking up prolonged sitting reduces postprandial glucose and insulin responses. *Diabetes Care*, Alexandria, v. 35, n. 5, p. 976-983, 2012.

DUVIVIER, B. M.; SCHAPER, N. C.; HESSELINK, M. K.; VAN KAN, L.; STIENEN, N.; WINKENS, B.; et al. Breaking sitting with light activities vs structured exercise: a randomised crossover study demonstrating benefits for glycaemic control and insulin

sensitivity in type 2 diabetes. *Diabetologia*, Berlin, v. 60, n. 3, p. 490-498, 2016.

ECKEL, R. H.; GRUNDY, S. M.; ZIMMET, P. Z. The metabolic syndrome. **The Lancet**, v. 365, n. 9468, p. 1415-1428, 2005. DOI: 10.1016/S0140-6736(05)66378-7.

EKTA, G.; TULIKA, M. G. Risk factor distribution for cardiovascular diseases among high school boys and girls of urban Dibrugarh, Assam. **Journal of Family Medicine and Primary Care**, v. 5, p. 108–113, 2016.

EMMANUEL, T.; MAUPONG, T.; MPOELENG, D.; SEMONG, T.; MPHAGO, B.; TABONA, O. A survey on missing data in machine learning. *Journal of Big Data*, v. 8, p. 1-37, 2021.

EVGENIOU, T.; PONTIL, M. Support vector machines: Theory and applications. In: **Machine Learning and Its Applications, Advanced Lectures**. September 2001, p. 249-257. DOI: 10.1007/3-540-44673-7_12.

FACCHINI, F. S. et al. Insulin resistance and cigarette smoking. **The Lancet**, v. 339, n. 8802, p. 1128-1130, 9 maio 1992.

FRIEDEWALD, W. T.; LEVY, R. I.; FREDRICKSON, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. **Clinical Chemistry**, v. 18, n. 6, p. 499-502, jun. 1972.

FRISCH, R. E.; REVELLE, R. The height and weight of girls and boys at the time of initiation of the adolescent growth spurt in height and weight and the relationship to menarche. *Human Biology*, Detroit, v. 43, n. 1, p. 140-159, 1971.

FROSIG, C. et al. Effects of endurance exercise training on insulin signaling in human skeletal muscle: Interactions at the level of phosphatidylinositol 3-kinase, Akt, and AS160. **Diabetes**, v. 56, n. 8, p. 2093-2102, ago. 2007.

GAMI, A. S. et al. Metabolic syndrome and risk of incident cardiovascular events and death: a systematic review and meta-analysis of longitudinal studies. **Journal of the American College of Cardiology**, v. 49, p. 403-414, 2007.

GERDS, Thomas A.; KATTAN, Michael W. **Medical Risk Prediction Models: With Ties to Machine Learning**. 1. ed. Chapman & Hall/CRC Biostatistics Series, 2022. ISBN 978-0367673734.

GONZAGA, M. T.; JOSÉ, A.; SOARES, F. O nível socioeconômico das escolas de educação básica brasileiras. 2012.

GOODFELLOW, I. et al. Deep Learning. **MIT Press**, 2016.

GOODMAN, E.; DANIELS, S. R.; DOLAN, L. M. Disparidades socioeconômicas na resistência à insulina: resultados do Estudo do Distrito Escolar de Princeton. **Psychosomatic Medicine**, v. 69, n. 1, p. 61-67, jan. 2007. DOI: 10.1097/01.psy.0000249732.96753.8f. Epub 2006 Dez 13. PMID: 17167128.

GORAN, M. I.; GOWER, B. A. Longitudinal study on pubertal insulin resistance. **Diabetes**, v. 50, n. 11, p. 2444-2450, 2001. doi:10.2337/diabetes.50.11.2444.

HAFFNER, S. M. et al. Prospective analysis of the insulin-resistance syndrome (syndrome X). **Diabetes**, v. 41, n. 6, p. 715-722, 1992.

HANNON, T. S.; JANOSKY, J.; ARSLANIAN, S. A. Longitudinal study of physiologic insulin resistance and metabolic changes of puberty. **Pediatr Res**, v. 60, n. 6, p. 759-763, dez. 2006. DOI: 10.1203/01.pdr.0000246097.73031.27. Epub 2006 Oct 25.

HERMAN-GIDDENS, M. E.; SLORA, E. J.; WASSERMAN, R. C.; BOURDONY, C. J.; BHAPKAR, M. V.; KOCH, G. G.; et al. Secondary sexual characteristics and menses in young girls seen in office practice: a study from the pediatric research in office settings network. *Pediatrics*, v. 99, n. 4, p. 505–512, 1997.

HEMMINGSSEN, B. et al. Diet, physical activity or both for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus. **Cochrane Database of Systematic Reviews**, 4 Dec. 2017.

HOWARD, B. V. et al. Relationships between insulin resistance and lipoproteins in nondiabetic African Americans, Hispanics, and non-Hispanic whites: the Insulin Resistance Atherosclerosis Study. **Metabolism: Clinical and Experimental**, v. 47, n. 10, p. 1174-1179, out. 1998.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Pesquisa Nacional da Saúde do Escolar (PeNSE) 2019. Rio de Janeiro: **IBGE**, 2019.

IWANI, A. K. N. Z. et al. TG: HDL-C ratio is a good marker to identify children affected

by obesity with increased cardiometabolic risk and insulin resistance. **International Journal of Endocrinology**, v. 2019, 2019.

JING, L.; TIAN, Y. Aprendizagem de características visuais auto-supervisionada com redes neurais profundas: uma pesquisa. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 11, p. 4037-4058, 2021.

KAHN, C. R. Insulin resistance, insulin insensitivity, and insulin unresponsiveness: a necessary distinction. **Metabolism: Clinical and Experimental**, v. 27, n. 12, suppl. 2, p. 1893-902, 1978.

KIM, D. S. Surrogate measures of insulin resistance in an apparently healthy population: a simpler and easier, yet Reliable Index. **Journal of Korean Medical Science**, v. 33, n. 29, p. e195, 2018. DOI: 10.3346/jkms.2018.33.e195.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. San Francisco: Morgan Kaufmann Publishers Inc., 1995. p. 1137-1143.

LEE, C. L.; LIU, W. J.; TSAI, S. F. Development and Validation of an Insulin Resistance Model for a Population with Chronic Kidney Disease Using a Machine Learning Approach. **Nutrients**, v. 14, n. 14, artigo 2832, jul. 2022. DOI: 10.3390/nu14142832.

LEVY-MARCHAL, C. et al. Insulin resistance in children: consensus, perspective, and future directions. **Journal of Clinical Endocrinology and Metabolism**, v. 95, n. 12, p. 5189-5198, 2010.

LI, C. et al. Prevalence of pre-diabetes and its association with clustering of cardiometabolic risk factors and hyperinsulinemia among U.S. adolescents: National Health and Nutrition Examination Survey 2005-2006. **Diabetes Care**, v. 32, n. 2, p. 342-347, 2009.

LI, G. et al. Effects of insulin resistance and insulin secretion on the efficacy of interventions to retard development of type 2 diabetes mellitus: The DA Qing IGT and Diabetes Study. **Diabetes Research and Clinical Practice**, v. 58, n. 3, p. 193-200, 2002.

LITTLE, R. J.; RUBIN, D. B. *Statistical Analysis with Missing Data*. v. 793, Hoboken: John Wiley & Sons, 2019. DOI: <https://doi.org/10.1002/9781119482260>.

LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, 2017.

MA, J. et al. Sugar-sweetened beverage but not diet soda consumption is positively associated with progression of insulin resistance and prediabetes. **The Journal of Nutrition**, v. 146, n. 12, p. 2544-2550, 1 dez. 2016.

MATTHEWS, D. R. et al. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. **Diabetologia**, v. 28, n. 7, p. 412-419, 1985.

MICHALISZYN, S. F.; ARSLANIAN, S. Techniques to assess insulin action in youth. In: ZEITLER, P. S.; NADEAU, K. J. (ed.). *Insulin resistance: childhood precursors of adult disease*. [S. l.]: **Springer Nature Switserzerland**, 2020.

MORAIS, P. R. S. et al. Correlation of insulin resistance with anthropometric measures and blood pressure in adolescents. **Arquivos Brasileiros de Cardiologia**, v. 106, n. 4, p. 319-326, abr. 2016. DOI: 10.5935/abc.20160041. Epub 2016 Mar 22.

MÜLLER, M. *Generalized Linear Models*. Beuth Hochschule für Technik Berlin, fevereiro 2004. DOI: 10.1007/978-3-642-21551-3_24.

MUST, A.; NAUMOVA, E. N.; PHILLIPS, S. M.; BLUM, M.; DAWSON-HUGHES, B.; RAND, W. M. Childhood overweight and maturational timing in the development of adult overweight and fatness: the Newton Girls Study and its follow-up. *Pediatrics*, v. 116, n. 3, p. 620-627, 2005. doi:10.1542/peds.2004-1604.

NATIONAL HIGH BLOOD PRESSURE EDUCATION PROGRAM WORKING GROUP ON HIGH BLOOD PRESSURE IN CHILDREN AND ADOLESCENTS. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. **Pediatrics**, v. 114, n. 2, suppl. 4th Report, p. 555-576, 2004

OGDEN, C. L.; CARROLL, M. D.; KIT, B. K.; FLEGAL, K. M. Prevalence of childhood and adult obesity in the United States, 2011–2012. *JAMA*, v. 311, n. 8, p. 806–814, 2014.

OIKONOMOU, E. K.; KHERA, R. Machine learning in precision diabetes care and cardiovascular risk prediction. **Cardiovascular Diabetology**, v. 22, p. 259, 2023. DOI: 10.1186/s12933-023-01985-3.

OWENS, S.; GALLOWAY, R. Childhood obesity and the metabolic syndrome. **Current Atherosclerosis Reports**, v. 16, n. 9, p. 436, 2014. DOI: 10.1007/s11883-014-0436-y.

OXFORD Centre for Evidence-Based Medicine. **Glossary of terms in Evidence-Based Medicine**. Disponível em: <http://www.cebm.net/glossary.asp>. Acesso em: 30 jun. 2024.

PANCH, T.; SZOLOVITS, P.; ATUN, R. Inteligência artificial, aprendizado de máquina e sistemas de saúde. **Journal of Global Health**, v. 8, n. 2, p. 020303, 2018.

PARK, S.; KIM, C.; WU, X. et al. Development and Validation of an Insulin Resistance Predicting Model Using a Machine-Learning Approach in a Population-Based Cohort in Korea. **Diagnostics (Basel)**, v. 12, n. 1, p. 212, 2022.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. Yandex, Moscow, Russia; **Moscow Institute of Physics and Technology**, Dolgoprudny, Russia, 2021. Disponível em: <https://doi.org/10.48550/arXiv.1706.09516>. Acesso em: 01 jul. 2024.

QUINLAN, J. R. Induction of Decision Trees. Machine Learning, Boston, v. 1, p. 81-106, 1986. RAMÍREZ-MANENT, J. I. et al. Waist Circumference Is an Essential Factor in Predicting Insulin Resistance and Early Detection of Metabolic Syndrome in Adults. **Nutrients**, v. 15, n. 2, artigo 257, jan. 2023. DOI: 10.3390/nu15020257.

REAVEN, G. M. Banting Lecture 1988. Role of insulin resistance in human disease. **Nutrition** (Burbank, Los Angeles County, Calif.), v. 13, n. 1, jan. 1997.

REINEHR, T.; WOLTERS, B.; KNOP, C.; LASS, N.; HOLL, R. W. Strong effect of pubertal status on metabolic health in obese children: a longitudinal study. *J Clin Endocrinol Metab*, v. 100, n. 1, p. 301-308, 2015. doi:10.1210/jc.2014-2674.

REN, L.; WANG, T.; SEKLOULI, A. S.; ZHANG, H.; BOURAS, A. A review on missing values for main challenges and methods. *Information Systems*, p. 102268, 2023.

REYES-SOFFER, G.; GINSBERG, H. N. Lipid and lipoprotein metabolism, hypolipidemic agents, and therapeutic goals. In: DEFRONZO, R. A.; ZIMMET, P.; ALBERTI, K. G. M. M. (Eds.). **International Textbook of Diabetes Mellitus**. John Wiley & Sons, Ltd, 2015.

RUDVIK, A.; MÅNSSON, M. Evaluation of surrogate measures of insulin sensitivity - Correlation with gold standard is not enough. **BMC Medical Research Methodology**, v. 18, n. 1, 2018.

SARKER, I. H. Aprendizado de máquina: algoritmos, aplicações no mundo real e direções de pesquisa. **SN Computer Science**, v. 2, n. 3, p. 160, 2021.

SATTAR, N.; WELSH, P.; PREISS, D. The insulin resistance syndrome. In: DEFRONZO, R. A. et al. (Eds.). **International Textbook of Diabetes Mellitus**. Chichester, UK: John Wiley & Sons, Ltd, 2015. p. 337-353.

SCHWARTZ, B. et al. Measurement of insulin sensitivity in children: comparison between the euglycemic-hyperinsulinemic clamp and surrogate measures. **Diabetes Care**, v. 31, n. 4, p. 783-788, 2008.

SHAHANI, N. M.; ZHENG, X.; GUO, X.; WEI, X. Machine learning-based intelligent prediction of elastic modulus of rocks at Thar Coalfield. **Sustainability**, v. 14, n. 6, art. 3689, mar. 2022. DOI: 10.3390/su14063689.

SIMONSON, D. C. Surrogate measures of insulin resistance: does one size fit all? **Diabetologia**, v. 58, n. 2, p. 207-210, 2014. DOI: 10.1007/s00125-014-3454-y.

SINGH, B.; SAXENA, A. Surrogate markers of insulin resistance: a review. **World Journal of Diabetes**, v. 1, n. 2, p. 36, 2010.

STERGIOU, G. S.; YIANNES, N. G.; RARRA, V. C. Validation of the Omron 705 IT oscillometric device for home blood pressure measurement in children and adolescents: The Arsakion School Study. **Blood Pressure Monitoring**, v. 11, n. 4, p. 229-234, 2006.

STUMVOLL, M. et al. Use of the oral glucose tolerance test to assess insulin release and insulin sensitivity. **Diabetes Care**, v. 23, n. 3, p. 295-301, 2000.

TANNER, J. M. Growth at Adolescence. 2. ed. **Oxford: Blackwell**, 1962.

TANNER, J. M.; WHITEHOUSE, R. H. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. *Archives of Disease in Childhood*, Londres, v. 51, n. 3, p. 170-179, mar. 1976. DOI: 10.1136/adc.51.3.170.

TELO, G. H. et al. Prevalence of type 2 diabetes among adolescents in Brazil: Findings from Study of Cardiovascular Risk in Adolescents (ERICA). ***Pediatric Diabetes***, v. 20, n. 4, p. 389-396, 2019.

TFAYLI, H.; ARSLANIAN, S. Pathophysiology of type 2 diabetes mellitus in youth: the evolving chameleon. ***Arquivos Brasileiros de Endocrinologia e Metabologia***, v. 53, n. 2, p. 165, mar. 2009.

THE R PROJECT FOR STATISTICAL COMPUTING. R: A Language and Environment for Statistical Computing [Internet]. Vienna: **R Foundation for Statistical Computing**, 2022 [citado em 9 abr. 2024]. Disponível em: <https://www.r-project.org/>

TIMMERMAN, K. L. et al. Altering physical activity influences insulin responses to glucose ingestion in healthy adults. ***International Journal of Sports Medicine***, v. 39, n. 13, p. 972-977, 2018. DOI: 10.1055/a-0735-9641.

UNITED NATIONS. Department of Economic and Social Affairs. Population Facts. New York: UN, 2019. Disponível em: https://www.un.org/en/development/desa/population/publications/pdf/popfacts/PopFacts_2019-6.pdf. Acesso em: 25 out. 2024.

VAN DER AA, M. P. et al. Population-Based Studies on the Epidemiology of Insulin Resistance in Children. ***Journal of Diabetes Research***, v. 2015, p. 1–9, 2015.

VASCONCELLOS, M. T. L. DE et al. Sampling design for the Study of Cardiovascular Risks in Adolescents (ERICA). ***Cadernos de Saúde Pública***, v. 31, n. 5, p. 921–930, maio 2015.

VICKERS, A. J.; VAN CALSTER, B.; STEYERBERG, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. ***BMJ***, v. 352, p. i6, 2016.

WAJCHENBERG, B. L.; LERARIO, A. C.; BETTI, R. T. B. Tratado de endocrinologia clínica. 2. ed. São Paulo: Grupo Editorial Nacional, 2014.

WALLACE, T.; LEVY, J.; MATTHEWS, D. Use and abuse of HOMA modeling. **Diabetes Care**, v. 27, n. 6, p. 1487-1495, 2004. DOI: 10.2337/diacare.27.6.1487.

WARREN, C. W. et al. Global youth tobacco surveillance, 2000-2007. **Morbidity and Mortality Weekly Report**, v. 57, n. 1, p. 1-28, 2008

WEISS, R. et al. Obesity and the metabolic syndrome in children and adolescents. **New England Journal of medicine**, v. 350, n. 23, p. 2362-2374, 2004.

WEITS, T.; VAN DER BEEK, E. J.; WEDEL, M.; TER HAAR ROMENY, B. M. Computed tomography measurement of abdominal fat deposition in relation to anthropometry. **International Journal of Obesity**, v. 12, n. 3, p. 217–225, 1988.

WORLD HEALTH ORGANIZATION. Global recommendations on physical activity for health. [S. l.]: WHO, 2010.

WORLD HEALTH ORGANIZATION. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. Geneva: WHO, 2020.

WORLD HEALTH ORGANIZATION. WHO guidelines on physical activity and sedentary behaviour. Geneva, 2020b. 93 p. Disponível em: <https://www.who.int/publications/i/item/9789240015128>. Acesso em: out. 2024.

WU, H. et al. Worldwide estimates of incidence of type 2 diabetes in children and adolescents in 2021. *Diabetes Research and Clinical Practice*, v. 185, p. 109785, 2022.

YALOW, R. S.; BERSON, S. A. Immunoassay of endogenous plasma insulin in man. **The Journal of Clinical Investigation**, v. 39, p. 1157-1175, 1960a.

YALOW, R. S.; BERSON, S. A. Plasma insulin concentrations in nondiabetic and early diabetic subjects. Determinations by a new sensitive immuno-assay technic. **Diabetes**, v. 9, p. 254-260, 1960b.

YARIBEYGI, H. et al. Molecular mechanisms by which aerobic exercise induces insulin sensitivity. **Journal of Cellular Physiology**, v. 234, n. 8, p. 12385-12392, 2019. DOI: 10.1002/jcp.28066.

YU, K. H.; BEAM, A. L.; KOHANE, I. S. Artificial intelligence in healthcare. **Nature**

Biomedical Engineering, v. 2, p. 719-731, 2018. DOI: 10.1038/s41551-018-0305-z.

ZAVARONI, I. et al. Hyperinsulinemia in a normal population as a predictor of non-insulin-dependent diabetes mellitus, hypertension, and coronary heart disease: the Barilla factory revisited. *Metabolism: Clinical and Experimental*, v. 48, n. 8, p. 989–994, ago. 1999.

ZEITLER, P. S.; NADEAU, K. J. (Eds.). *Insulin resistance: childhood precursors of adult disease*. [S. l.]: **Springer Nature Switzerland**, 2020.

ZHANG, H. et al. Development and validation of machine learning-augmented algorithm for insulin sensitivity assessment in the community and primary care settings: a population-based study in China. *Frontiers in Endocrinology* (Lausanne), v. 15, artigo 1292346, 25 jan. 2024. DOI: 10.3389/fendo.2024.1292346.

ZHANG, Q.; WAN, N. J. Simple Method to Predict Insulin Resistance in Children Aged 6–12 Years by Using Machine Learning. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, v. 15, p. 2963-2975, 2022. DOI: 10.2147/DMSO.S380772.

ZIMMET, P.; ALBERTI, K. G.; SHAW, J. Global and societal implications of the diabetes epidemic. *Nature*, v. 414, n. 6865, p. 782–787, 2001.