ESCOLA SUPERIOR DE CIÊNCIAS DA SANTA CASA DE MISERICORDIA DE VITÓRIA - EMESCAM

BRUNA NASCIMENTO ARRUDA SCABELLO

GABRIEL GOMES PEREIRA DE AGUIAR PEROBA

VICTOR PEYNEAU PONCIO

RISCO CARDIOVASCULAR EM ADOLESCENTES: UMA APLICAÇÃO DO MACHINE LEARNING NA INTERPRETAÇÃO DE DADOS

VITÓRIA
2021

BRUNA NASCIMENTO ARRUDA SCABELLO
GABRIEL GOMES PEREIRA DE AGUIAR PEROBA
VICTOR PEYNEAU PONCIO

RISCO CARDIOVASCULAR EM ADOLESCENTES: UMA APLICAÇÃO DO MACHINE LEARNING NA INTERPRETAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado à Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM, como requisito parcial para obtenção do grau de médico.

Orientadora: Profª. Dra. Katia Valéria Manhabusque
Coorientador: Prof. Dr. Gustavo Carreiro Pinasco

VITÓRIA
2021

BRUNA NASCIMENTO ARRUDA SCABELLO
GABRIEL GOMES PEREIRA DE AGUIAR PEROBA
VICTOR PEYNEAU PONCIO

RISCO CARDIOVASCULAR EM ADOLESCENTES: UMA APLICAÇÃO DO MACHINE LEARNING NA INTERPRETAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de Medicina da Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM, como requisito parcial para obtenção do grau de médico.

Aprovado em 20 de maio de 2021

BANCA EXAMINADORA

Profª. Dra. Katia Valéria Manhabusque
Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM
Orientadora

Prof. Dr. Gustavo Carreiro Pinasco
Universidade Federal do Espírito Santo - UFES
Coorientador

Profª. Dra. Patrícia Casagrande Dias De Almeida
Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM
Avaliadora

Agradecemos ao coorientador Dr. Gustavo Carreiro Pinasco pela oportunidade do crescimento e aprendizado científico, ao nos dar a oportunidade de transitar por diferentes campos do conhecimento, percebendo, assim, a importância da formação e conhecimento integral do profissional.

RESUMO

Objetivo: O estudo pretende avaliar a capacidade de predizer por meio do machine learning alterações de exames laboratoriais relacionados ao aumento de risco cardiovascular. **Método:** Consiste em um estudo transversal, sendo usado como população a referida pelo estudo “Excessive weight, cardiovascular risk and metabolic syndrome in adolescents from public educational system of Metropolitan Region of Grande Vitória, Brazil”. Foram obtidos dados antropométricos e laboratoriais de 817 adolescentes de ambo sexos. A análise de dados foi realizada com técnicas de machine learning, sendo utilizada a metodologia *Classic Shapley Value Estimation* (SHAP) para facilitar a interpretabilidade do modelo construído. **Resultado:** Observou-se que as dobras cutâneas tricipital e subescapular demonstraram maior impacto sobre os modelos preditivos de níveis de triglicerídeos, de colesterol total e VLDL, os níveis sanguíneos de glicose, HDL e LDL colesterol tiveram maior poder de previsão evidenciada pela circunferência do pescoço e a circunferência da cintura teve seu impacto mais relacionado aos valores de insulina e de PCR. **Conclusão:** Os dados e o projeto ressaltam que a análise preditiva pelo método de machine learning com aplicação na saúde apresenta-se como um grande potencial para identificar relações complexas e não-lineares presentes nos dados.

Palavras-chave: Aprendizado de máquina. Doenças Cardiovasculares. Circunferência do pescoço.

ABSTRACT

Objective: The study proposes to evaluate the ability to predict changes in laboratory tests related to increased cardiovascular risk through machine learning. **Method:** It consists of a cross-sectional study, the population used is the one referred to the study “Excessive weight, cardiovascular risk and metabolic syndrome in adolescents from the public educational system of Metropolitan Region of Grande Vitória, Brazil”. Anthropometric and laboratory data were obtained from 817 adolescents of both genres. The data analysis was performed with machine learning techniques, using the Classic Shapley Value Estimation (SHAP) methodology to facilitate the interpretability of the constructed model. **Results:** It was observed that the tricipital and subscapular skinfolds demonstrated a greater impact on the predictive models of triglyceride levels, total cholesterol and VLDL, blood glucose, HDL and LDL cholesterol levels had greater predictive power evidenced by the circumference of the neck and the waist circumference had its impact more related to insulin and CRP values. **Conclusion:** The data and the project emphasize that predictive analysis by the machine learning method with application in health is a great potential to identify complex and non-linear relationships present in the data.

Keyword: Cardiovascular diseases. Machine learning. Neck circumference.

LISTA DE ILUSTRAÇÕES

Página

LISTA DE TABELAS

Página

LISTA DE GRÁFICOS

Página

LISTA DE ABREVIATURAS

SHAP - C*lassic Shapley Value Estimation*

LISTA DE SIGLAS

C Braço – Circunferência braquial
CC – Circunferência da cintura
C Pescoço – Circunferência do pescoço
CP – Circunferência do pescoço
CSV – Central Sorológica de Vitória
CT – Colesterol total
DCSE – Dobra cutânea subescapular
DCT – Dobra cutânea tricipital
DP – Desvio padrão
EUA – Estados Unidos da América
HDL-C – Lipoproteína de colesterol de alta densidade
IMC – Índice de massa corporal
LDL-C – Lipoproteína de colesterol de baixa densidade
OMS – Organização Mundial da Saúde
PA – Pressão arterial
PCR-us – Proteína C reativa ultrassensível
RMSE – Erro quadrático médio
TG – Triglicérides
VLDL-c – Lipoproteína de colesterol de muito baixa densidade

SUMÁRIO

1 INTRODUÇÃO

As doenças cardiovasculares apresentam expressiva prevalência na população, e segundo a Organização Mundial da Saúde (OMS), constituem a principal causa de óbito no mundo e no Brasil.[1-3] Pelas informações da Sociedade Brasileira de Cardiologia, foram 383.961 mortes em 2017, com estimativa de que ao final de 2021, quase 400 mil cidadãos brasileiros morrerão por doenças do coração e da circulação.[4] Sabe-se que os métodos disponíveis para a estratificação de risco cardiovascular levam em conta, entre outras medidas, fatores de risco modificáveis no indivíduo que conduzem a um desfecho desfavorável. Estudos feitos a longo prazo, evidenciam como deve ser prioritária a prevenção em saúde a fim de alterar a morbimortalidade destas doenças.[5] Portanto, quanto mais precoce a identificação desses fatores, maior a oportunidade para aplicar estratégias de intervenção.[6]

Diante deste contexto, fica evidente a importância de dispormos de modelos preditivos que sejam capazes de mostrar, através de medidas simples, adolescentes que apresentem maior potencial a terem complicações cardiovasculares na vida adulta.[7,8] Uma das formas de se executar esta tarefa é através de modelos computacionais programados, pelo chamado *machine learning*, que utiliza algoritmos para analisar em pouquíssimo tempo um vasto número de variáveis provenientes de uma base populacional, procurando combinações que preveem resultados com confiabilidade, e ainda permitindo simultaneamente estabelecer uma correlação entre eles.[9,10]

Neste estudo, partindo-se de uma população de ambos os sexos, da faixa etária de 10 até 14 anos, foi feita uma comparação entre medidas antropométricas e laboratoriais, para que fosse possível interligá-las àquelas que representam fatores de risco das doenças cardiovasculares. Deste modo, foi possível dizer qual medida é capaz de prever aumento do risco. Foram avaliadas: pressão arterial (PA), estatura e peso para cálculo do índice de massa corporal (IMC), circunferência da cintura (CC), dobras cutâneas tricipital (DCT) e subescapular (DCSE), circunferência de pescoço (CP), glicose sérica, insulina sérica, colesterol total (CT), Lipoproteína de colesterol de alta densidade (HDL-C), lipoproteína de colesterol de baixa densidade (LDL-C),

Lipoproteína de colesterol de muito baixa densidade (VLDL-C), triglicérides (TG) e proteína C reativa ultrassensível (PCR-us).

Por fim, é válido ressaltar que o *machine learning* criado neste estudo serve como exemplo do potencial do método, e abre caminho para que outros possam utilizar em larga escala esta estratégia rápida e segura, a fim de programar intervenções que possam mudar o cenário da saúde a curto e longo prazo. Portanto, tem-se como principal objetivo avaliar o impacto e construção de um modelo preditivo que correlaciona variáveis laboratoriais e antropométricas evidenciando o risco de doenças cardiovasculares em adolescentes.

2 MÉTODO

2.1 Desenho do Estudo e População Analisada

Para o estudo transversal em questão, a população escolhida foi baseada no estudo "*Excessive weight, cardiovascular risk and metabolic syndrome in adolescents from public educational system of Metropolitan Region of Grande Vitória, Brazil*". Os dados foram coletados entre agosto de 2012 e outubro de 2013, sendo selecionadas adolescentes com idade de 10 a 14 anos (idade média = 12,83 anos ± 1,14 de desvio padrão) por meio da randomização de 817 de adolescentes de ambos os sexos (meninos = 340; meninas = 477), das escolas públicas da região da Grande Vitória, Espírito Santo, Brasil. Houve coleta da população em 7 cidades diferentes, as quais compõem a região metropolitana com cerca de 3,5 milhões de habitantes. Para isso, utilizou-se uma equipe devidamente treinada para realização das medições de dados laboratoriais e antropométricos e registro do obtido.

A amostra foi calculada pela fórmula de população infinita. Considerou-se um nível de confiança de 95%, erro de amostragem de 3%. Logo, a amostra estimada foi de 667 adolescentes. Contudo, para a coleta estimou-se uma perda amostral de 30%, que não se confirmou, tendo-se chegado ao número final de 817 adolescentes avaliados. Porém diante da impossibilidade da dosagem laboratorial de alguns adolescentes avançaram para a próxima fase do modelo 699 variáveis.

Os adolescentes que apresentavam história de doenças inflamatórias agudas e crônicas, obesidade secundária, uso de corticosteroides e/ou anti-inflamatórios foram excluídas do estudo.

O estudo foi realizado conforme a Declaração de Helsinque e atendendo as especificações do comitê de ética e pesquisa número 466/12. Procedimentos envolvendo voluntários humanos obtiveram aprovação no comitê de ética institucional. Aos adolescentes e seus pais ou responsáveis legais foi requerida a assinatura do Termos de Consentimento Livre e Esclarecido para participação no estudo.

2.2 Antropometria

Os dados antropométricos dos pacientes selecionados para o estudo foram obtidos a partir de equipamentos aferidos e validados. Os adolescentes foram pesados descalços e usando roupas leves em balanças eletrônicas portáteis Tanita® A-080 (Arlington Heights, Illinois, EUA), com capacidade máxima de 150 Kg e graduação de 0,1 Kg. Para o comprimento foi utilizado o estadiômetro telescópio portátil Alturaexata® (Belo Horizonte, Minas Gerais, Brasil); foi utilizado um intervalo de 0,001 - 214 centímetros. Com base nas medidas de peso e estatura, calculou-se o índice de massa corporal (IMC) seguindo a fórmula “IMC = Peso (Kg)/Estatura (m)2”. Para a classificação quanto aos estados nutricionais e aos procedimentos para a obtenção das medidas antropométricas foram utilizadas as recomendações da Organização Mundial da Saúde (OMS).[11, 12]

A circunferência da cintura foi medida por meio de uma fita milimetrada, da marca Sanny ®, de máxima extensão de 200 cm e precisão de 0,1 cm, sendo colocado ao nível da cicatriz umbilical e classificado para ponto de corte o percentil 90, segundo idade e sexo.[13]

As dobras cutâneas, foram medidas por meio de um compasso de dobras nos seguintes pontos: subescapular, tríceps, bíceps, peitoral, axilar média, supra ilíaca, abdominal, coxa e panturrilha medial. Sendo utilizado para o presente estudo principalmente as medidas subescapular e tricipital.

A circunferência de pescoço foi medida com o paciente em posição ortostática ao nível da cartilagem tireoide (até 0,1 cm mais próximo), com fita inelástica e com a cabeça mantida em posição ereta seguindo recomendações científicas.[14]

2.3 Laboratório

As variáveis laboratoriais foram analisadas por meio da coleta de 10 ml de sangue, em 12 horas de jejum, seguindo adequadamente técnicas assépticas com materiais descartáveis que foram propriamente identificados. A coleta ocorreu no próprio local de estudo dos participantes, durante o período matutino, por profissionais qualificados e legalmente habilitados. Os níveis séricos de glicose foram dosados e analisados

pelo laboratório da Centse; insulina, colesterol total, HDL-C (Lipoproteína de colesterol de alta densidade), LDL-C (Lipoproteína de colesterol de baixa densidade), VLDL-C (Lipoproteína de colesterol de muito baixa densidade), triglicérides (TG) e de PCR-us (Proteína C reativa ultrassensível) pelo Central Sorológica de Vitória (CSV).

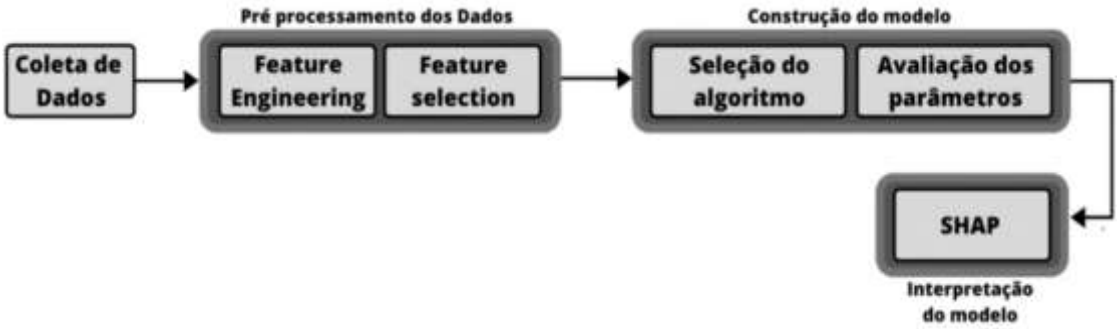
2.4 Análise dos dados

A análise de dados foi realizada com técnicas de *machine learning* seguindo um fluxo de boas práticas na área, como exemplificado na Figura 1. O desenvolvimento do algoritmo começou pelo pré-processamento de dados, um conjunto de atividades que envolvem preparação, organização e estruturação dos dados, ponto de grande importância, pois será determinante para a qualidade final dos dados que serão analisados. No estudo em questão foi dividido em duas etapas a fins didáticos: *feature engineering* e a *feature selection*, processos que serão mais bem explicados em seguida.

Foi criado um modelo de regressão, de aprendizado supervisionado, ou seja, tinha-se conhecimento dos dados fornecidos e as saídas esperadas, contudo era necessário que se passasse para o sistema de aprendizagem, que visa avaliar melhores correlações e caminhos ajustando o próprio modelo para chegar aos resultados esperados.

Após esse processo, se fez a validação do modelo com a divisão dos dados pré-processados em treino e teste, sendo testada a performance de algoritmos que foram rankeados, permanecendo o de melhor performance para a construção do modelo final. De modo a evitar que os algoritmos apresentem resultados não facilmente interpretável, o efeito *black box*, utilizou-se da metodologia *Classic Shapley Value Estimation* (SHAP) para tornar a interpretabilidade do modelo mais clara e mensurar a contribuição de cada variável preditora dentro do modelo desenvolvido.

Figura 1 – Roteiro de realização de *machine learning*



Fonte: Elaboração própria, 2021

2.4.1 *Feature Engineering*

É relevante que essa etapa não seja ignorada, pois a qualidade dos dados inseridos possui um protagonismo para criar ou quebrar a capacidade preditiva de um modelo. Dentro dessa lógica, deve-se ter em mente que diferentes modelos têm sensibilidades distintas aos tipos de preditores utilizados e o método de seleção dos preditores para ingressar no modelo também não pode ser desconsiderado.

No estudo em questão, a partir dos dados do estudo já citado anteriormente foi criado uma base de dados estruturados, em modelo excel (*.csv*), foi então avaliada cada coluna individualmente, ou seja, avaliado as *features* mais relevantes, e realizado ajustes de modo que as variáveis que estavam em formatos não compreendidos ou que pudessem comprometer a leitura do modelo foram ajustadas para um formato padrão dentro da linguagem *Python*.

Outra parte importante dessa fase foi a busca por valores faltantes ou *missing*, pois estes podem produzir um viés significativo no modelo. Primeiramente, é importante buscar quais valores estão ausentes, observar se há um padrão de dados ausentes. No caso do modelo em questão, os dados faltantes se referiam a exames laboratoriais, isso é explicado pelo fato de que nem todas os adolescentes examinados foram fazer exames laboratoriais. As variáveis que apresentaram dados faltantes foram GLICOSE (mg/dL), GLICOSE (mmol/L), INSULINA (mcU/mL), TG (mg/dL), COLESTEROL T(mg/dL), HDL (mg/dL), LDL (mg/dL), VLDL (mg/dL), PCR-US (mg/L), totalizando 118 variáveis preditoras. O método usado para o tratamento desses dados foi o descarte das linhas que continham esses valores constantes, sendo que esta ação não altera

o impacto do modelo final. Dessa forma, avançaram para a próxima fase do modelo 699 variáveis preditoras.

2.4.2 *Feature Selection*

Com a recategorização dos dados, avançou para a próxima etapa dentro do *Feature Engineering*: selecionar o número de *features*, tanto quantitativo, quanto qualitativamente. Aqui é explícita a notoriedade do cientista de dados compreender a relevância de cada *feature*, pois a avaliação dos modelos não se dá apenas pelos valores, contudo na resposta que ele pode trazer aos problemas propostos, sendo fundamental fornecer ao algoritmo as *features* mais relevantes. De outro modo, com muitas *features*, ou com *features* irrelevantes, o modelo será pouco expressivo.

Diante dessa etapa, é fundamental compreender que cada f*eature* possui dispersões próprias. Assim, assumindo que ela será comparada pelos algoritmos, f*eatures* com maior dispersão tenderão a dominar o modelo, o tornando ineficaz, com baixa reprodutibilidade. Dessa forma, é necessário que os dados passem por um processo chamado de padronização ou normalização do escore Z, onde eles serão centralizados e redimensionados. Para o modelo preditivo utilizado, a técnica mais adequada foi o *Standard Scaler*, em que ocorre uma subtração da média da variável (em todos os pontos de dados) e divisão pela variância, de forma a normalizar a distribuição dos dados.

2.4.3 Construções dos Modelos

Feito o pré-processamento, a próxima etapa consistiu na construção dos modelos, sendo utilizada uma biblioteca de machine learning de código aberto para a linguagem de programação *Python*, o *Scikit-Learn.* Primeiramente realizou-se a divisão aleatória do conjunto de dados, em treinamento e teste com a ajuda do método *train_test_split* da biblioteca do *Scikit-Learn*, sendo feita a separação do banco de dados em 70% (Treino) e 30% (Teste). É importante ressaltar que nas etapas de construção e treino do modelo apenas o conjunto de treinamento é explorado, assegurando ao final do processo, a obtenção de resultados com maior performance preditiva quando aplicado a novas observações.[15,16]

Buscou-se, inicialmente, o ajuste adequado do modelo utilizado para dados tabulares. Foi usado como modelo preditivo o *RandomForest*, que permite ajustar uma série de árvores de decisão em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e o controle de sobreajuste.

Para avaliação da performance do modelo preditivo foi definido como a melhor métrica a raiz quadrada do erro quadrático médio (*RMSE - Root Mean Squared Error*), que consiste na raiz quadrada da diferença quadrática média entre os valores estimados e o que é obtido como resultado, evidenciando o quão espalhados estão os dados. Em conjunto com o parâmetro acima, também foi observado o coeficiente R^2, chamado por vezes de coeficiente de determinação, que permite avaliar quão bem os resultados reais são replicados pelo modelo regressor, baseando-se na variação total da previsão explicada pelo modelo. O valor do coeficiente tem variação entre 0 e 1 (0% e 100%), porém pode ser negativa, podendo indicar um modelo arbitrariamente pior, fato importante a ser descrito é que um banco de dados com quantidade de dados pouco expressiva para o modelo pode levar a um o coeficiente de baixa estatisticamente significativa, ou seja, pouco expressivo para interpretação no caso.[17]

2.4.4 Explicação dos Modelos

Esta etapa mostra-se como diferencial e torna-se cada vez mais impactante, pois o uso de um método que aumenta a transparência do modelo, permitindo o compromisso entre precisão e interpretabilidade impacta diretamente no uso do machine learning na prática. Para o modelo usado neste estudo foi escolhido a explicação pela metodologia do Shapley Additive Explanations (SHAP) por discorrer da importância das variáveis preditoras tanto globalmente quanto localmente.

A escolha pelo SHAP deu-se devido a capacidade de interpretar globalmente o modelo, evidenciando o quanto cada preditor contribui, positiva ou negativamente, para a variável de destino. Outro fator relevante é a capacidade de interpretabilidade local, ou seja, cada observação obtém seu próprio conjunto de valores SHAP, permitindo apontar e constatar os impactos dos fatores, aumentando sua transparência. É importante ressaltar, nesse contexto, que os valores de SHAP não fornecem causalidade.[18]

3 RESULTADO

A caracterização da amostra dos 817 adolescentes avaliados está demonstrada na Tabela 1. Desses 58,38 % são do sexo feminino e 41,62 % do sexo masculino. A média de idade em meses é 154.06 (13.75), escore z de IMC 0.20 (1.22), circunferência do pescoço 30.71 (2.41). Foi apresentado as variáveis bioquímicas, o qual não observou valores médios alterados na amostra, sendo o valor médio do colesterol total (162.53 ± 28.90) apresentado valor mais próximo do limite superior de normalidade.

Tabela 1 - Descrição de características clínicas e laboratoriais dos adolescentes.

Variáveis	**Média (DP)**
Idade (meses)	154.06 (13.75)
Variáveis antropométricas	
Peso (Kg)	48.51 (11.77)
Estatura (cm)	155.50 (8.90)
IMC (Kg/m²)	19.87 (3.74)
Circunferência da cintura (cm)	70.41 (9.74)
Circunferência Pescoço (cm)	30.71 (2.41)
Dobra Cutânea Tricipital	15.15 (6.13)
Dobra Cutânea Subescapular	11.27 (6.10)
Variáveis bioquímicas	
Glicose (mg/dl)	86.42 (8.25)
Insulina (mcu/ml)	13.34 (8.90)
Triglicerídeos (mg/dl)	80.77 (39.12)
Colesterol t (mg/dl)	162.53 (28.90)
HDL (mg/dl)	53.36 (15.90)
LDL (mg/dl)	98.17 (24.48)
VLDL (mg/dl)	15.89 (8.91)
PCR (mg/l)	1.21 (3.34)
Variáveis	**n (%)**
Sexo (meses)	
Feminino	477 (58.38)
Masculino	340 (41.62)

Fonte: Elaboração própria, 2021

A tabela 2 evidencia os parâmetros utilizados para avaliação da performance do modelo preditivo. O RMSE apresentou valores menores nos parâmetros PCR, insulina, glicose e VLDL, respectivamente, 3,83; 6,86; 8,22; 8,46, evidenciando esses como os menores erros reais diante do proposto. Outro parâmetro utilizado foi o coeficiente R^2, que permitiu identificar o modelo preditivo de insulina com índice de

0,1446 como o melhor quando comparado aos outros modelos propostos no trabalho. Nota-se que os demais modelos construídos apresentaram performance negativa, fato que não exclui a possibilidade de avaliação, porém ressalta que um banco de dados com dados quantitativamente pouco expressivos compromete a avaliação individual e global dos modelos preditivos.

Tabela 2 - Parâmetros de avaliação de performance do modelo preditivo

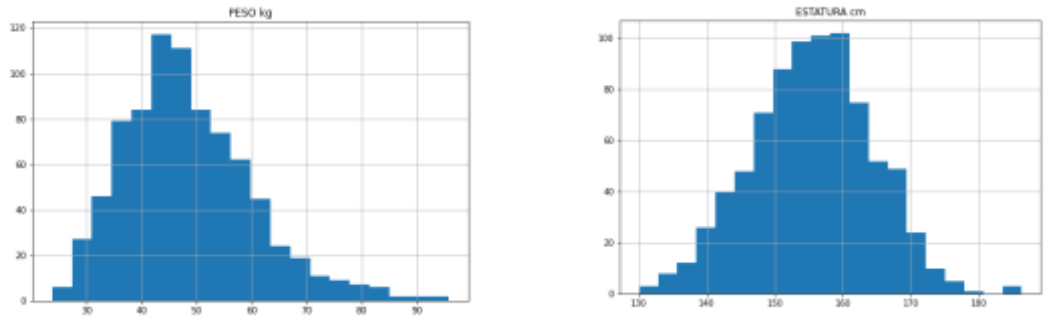
Variáveis	RMSE	R^2
Glicose	8,2227	- 0,1148
Insulina	6,8605	0,1446
Triglicerídeo	41,0998	- 0,2247
Colesterol Total	28,5475	- 0,1768
HDL	51,1093	- 0,0759
LDL	25,6515	- 0,1051
VLDL	8,4609	- 0,1862
PCR	3,8327	- 0,0213

Fonte: Elaboração própria, 2021

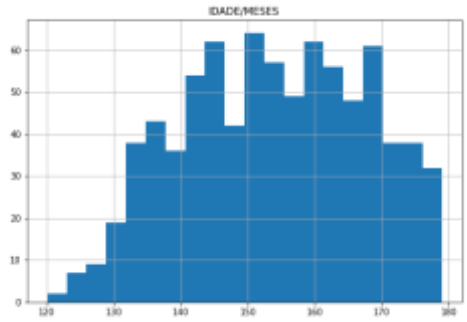
A dispersão dos valores de algumas das variáveis pode ser observada no Gráfico 1. Percebe-se que em alguns casos, como nos gráficos de dispersão de estatura e colesterol, ocorre uma dispersão mais uniforme. Enquanto outros apresentam dispersões mais dispersas e centradas em extremidades, em alguns casos podem ser explicado pelos baixos valores como no caso do PCR-us e de Insulinas. Importante ressaltar que para aplicação de modelos preditivos foi realizado a padronização ou normalização dessas dispersões de maneira que não haja dominação do modelo por uma única variável.

Gráfico 1 – Gráficos de dispersão das variáveis **a.** Peso **b.** Estatura **c.** Idade **d.** PCR-us **e.** Glicose **f.** Insulina **g.** Colesterol **h.** Triglicérides

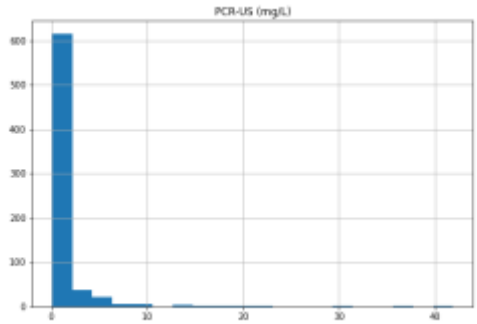
a. **b.**



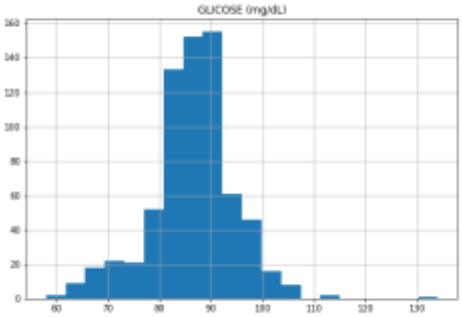
c.



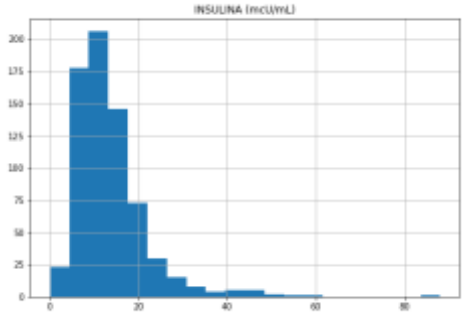
d.



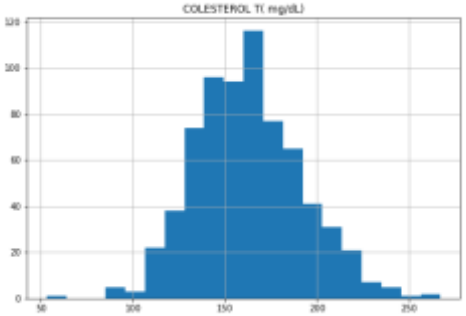
e.



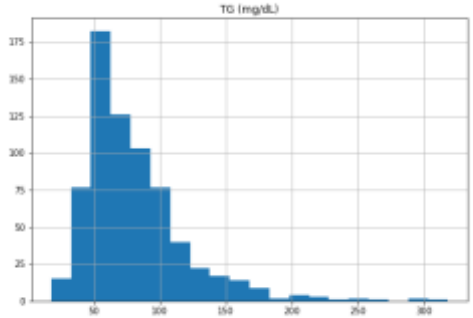
f.



g.



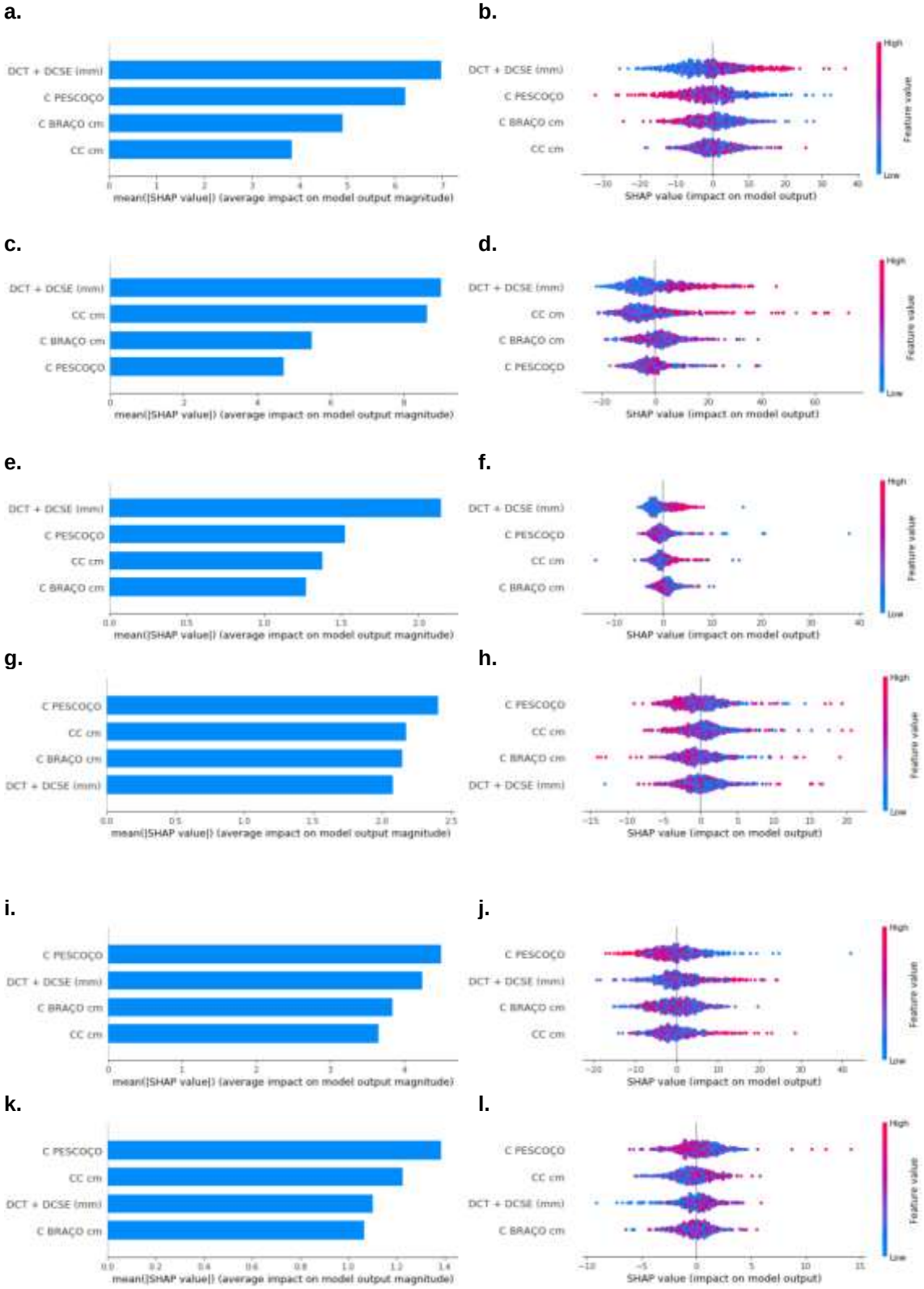
h.



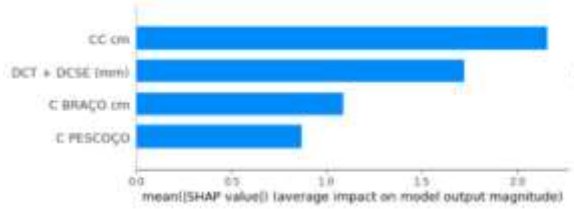
Fonte: Elaboração própria, 2021

A interpretabilidade dos modelos preditivos de cada variável antropométrica em relação às variáveis laboratoriais está demonstrada no Gráfico 2. Os gráficos de resumo (“*Summary plots*”) são usados para evidenciar quais variáveis de maior impacto no modelo. Primeiramente, observou-se que as dobras cutâneas tricipital e subescapular demonstraram maior impacto sobre os modelos preditivos de níveis de triglicerídeos, de colesterol total e VLDL. Em segundo lugar, os níveis sanguíneos de glicose, HDL e LDL colesterol tiveram maior poder de previsão evidenciada pela circunferência do pescoço. Por fim, a circunferência da cintura teve seu impacto mais relacionado aos valores de insulina e de PCR.

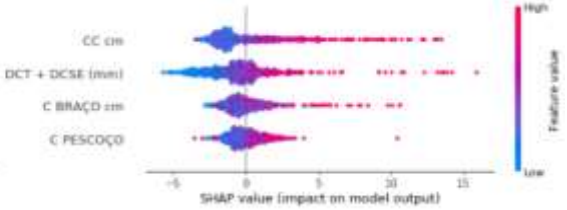
Gráfico 2 – Gráfico resumo do valor *Shap* para os modelos preditivos das variáveis laboratoriais. **a. e b.** modelo preditivo de colesterol total; **c. e d.** modelo preditivo de triglicerídeos; **e. e f.** modelo preditivo de VLDL; **g. e h.** modelo preditivo de HDL; **i. e j.** modelo preditivo de LDL; **k. e l.** modelo preditivo de glicose; **m. e n.** modelo preditivo de insulina; **o. e p.** modelo preditivo de PCR.



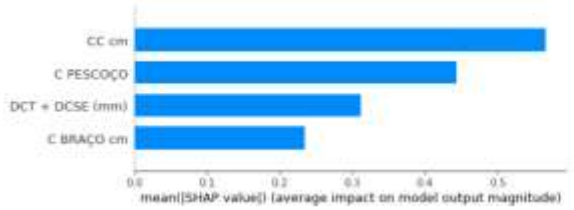
m.



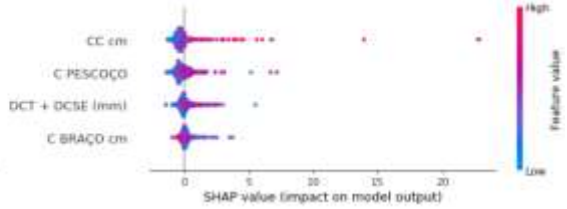
n.



o.



p.



Fonte: Elaboração própria, 2021

4 DISCUSSÃO

Nesse estudo foi proposto avaliar o impacto entre diferentes parâmetros antropométricos com amostras laboratoriais dos adolescentes participantes do estudo anteriormente citado. Estudos anteriores avaliaram a associação entre medidas antropométricas e parâmetros metabólicos de forma estatística linear, usando em maioria das vezes modelos de regressão linear, contudo o presente estudo traz a construção de modelos preditivos que permitem avaliar o impacto de cada variável antropométrica sobre os parâmetros laboratoriais, procurando explicar as variáveis de forma não linear por meio do *machine learning*.

Foi observado para o modelo preditivo construído para avaliar a glicose plasmática livre que a circunferência de pescoço obteve maior impacto na predição quando comparado a outras medidas. A literatura reforça o que foi evidenciado pelos dados, demonstrado que a correlação da circunferência do pescoço se mantém maior do que as demais, sendo inferior apenas a circunferência do quadril, medida antropométrica não avaliada no presente estudo.[19] Nela também observa-se correlação positiva entre a glicose plasmática e a circunferência do pescoço.[20] A mesma correlação foi observada entre a circunferência de cintura e a glicose, entretanto nas análises de modelo de predição a circunferência de cintura apresentou-se como uma variável com menor capacidade de explicação do resultado preditivo em comparação com a circunferência do pescoço.

Houve maior impacto da circunferência de cintura sobre o modelo da insulina plasmática, o que, por sua vez, não acompanhou o que foi visto para a glicose. Apesar de os níveis de insulina e glicose se relacionarem quando se trata da fisiologia humana, o que pode ser observado no presente estudo é a diferença de impacto das medidas antropométricas sobre os modelos preditivos das referidas medidas laboratoriais. Existem artigos que vão de encontro ao achado do presente estudo, quando evidencia que a circunferência de pescoço foi o marcador mais consistente e robusto ao se correlacionar com valores de insulina.[21]

Também através do modelo de predição de impacto, constatou-se que os modelos criados para observar triglicerídeos, colesterol total e VLDL foram mais fortemente

impactados pelas dobras cutâneas tricipital e dobras cutâneas subescapulares.[22, 23] Em trabalhos prévios a este, também constatou-se significância entre as variáveis em questão, com forte associação especialmente entre os triglicerídeos e as dobras cutâneas nos adolescentes, indo de acordo com o presente trabalho, e inclusive encontrando valores tão significativos quanto quando a associação foi feita em relação ao IMC, parâmetro mais usualmente difundido nos estudos antropométricos.[24] O mesmo também acontece em outros. [25] Em estudos mais recentes, a dobra cutânea subescapular demonstra tanta precisão quanto o IMC e a circunferência de cintura para predição (rastreio epidemiológico) do risco cardiometabólico em crianças e adolescentes, portanto sua utilização deve ser encorajada.

Quando se observa, no entanto, os modelos de HDL e LDL, a circunferência de pescoço mostra-se como fator de maior impacto. Alguns autores demonstraram resultados que vão de encontro aos aqui expostos, evidenciando uma correlação negativa entre a circunferência de pescoço e HDL em adolescentes de ambos os sexos e em estágios pré-puberal e puberal.[26] Outros apresentaram o HDL e o LDL se associaram positivamente com a circunferência de pescoço, porém o último apenas em meninos; neste estudo também pode-se notar que houve associação positiva dos índices de colesterol citados anteriormente com a circunferência de pescoço do que com a circunferência de cintura, corroborando com o que pode ser observados nos modelos preditivos propostos.[27] Uma meta-análise produzida em 2018 enfatiza uma correlação negativa entre HDL e circunferência do pescoço em relevantes publicações, porém não sendo um consenso entre todas, por outro lado foi evidenciado correlações fracas e pouco expressivas entre LDL e circunferência do pescoço.[28]

O presente estudo avaliou a relação positiva entre circunferência de cintura e PCR, dados que também são associados entre si em outra referência, sendo, segundo o autor, o primeiro a estabelecer pontos de corte para as medidas de adiposidade central a fim de identificar adolescentes com alto risco de ter fatores como a proteína C reativa elevados.[29]

5 LIMITAÇÕES DO ESTUDO

A interpretação global e individual dos modelos preditivos pelo SHAP *value* se beneficia de um maior número de dados e este estudo pode ter apresentado um número de indivíduos reduzido, o “n” pode ter influenciado o resultado deste teste em algumas análises. Outra limitação observada foi a presença de dados faltantes em algumas variáveis devido a não realização de exames delas nos participantes do estudo.

Em contrapartida há uma grande quantidade de variáveis analisadas, o que torna a aplicação do modelo regressor de machine learning válido assim como sua interpretabilidade e avaliação dos resultados.

6 CONCLUSÃO

A análise preditiva pelo método de *machine learning* com aplicação na saúde apresenta-se como um grande potencial para identificar relações complexas e não-lineares presentes nos dados. A prática clínica é rica em dados que quando correlacionados entre si permitem a extração de informações relevantes para comunidade científica. O presente estudo evidencia o impacto relevante dos modelos preditivos sobre a análise de fatores de grande impacto no risco cardiovascular de adolescentes. Foi demonstrado que a intercambialidade entre diferentes áreas do conhecimento apenas contribui para a robustez das informações, o que auxilia a evolução da prática médica. A utilização para manejo clínico e para o respaldo na tomada de decisões requer estudos adicionais que avaliem esta ferramenta em outros cenários e outras faixas de população pediátrica.

REFERÊNCIAS

1. World Health Organization. (WHO) Global Action Plan for the Prevention and Control of NCDs 2013-2020. Geneva; 2013

2. Marinho FM, Passos V, Malta DC, Barbosa FE, Abreu DMX. Burden of disease in Brazil, 1990-2016: a systematic subnational analysis for the Global Burden of Disease Study 2016. Lancet. 2018 Sep 1;392(10149):760-75.

3. Schmidt MI, Duncan BB, Azevedo e Silva G, Menezes AM, Monteiro CA, et al. Chronic noncommunicable diseases in Brazil: burden and current challenges. Lancet. 2011; 377(9781):1949-61.

4. Sociedade Brasileira De Cardiologia Na Rede [Internet]. Rio de Janeiro: SBC; c2015 [cited 2021 Mar 19]. Available from: http://www.cardiometro.com.br/anteriores.asp

5. Ciorlia LAS, Godoy MF. Fatores de risco cardiovascular e mortalidade: seguimento em longo prazo (até 20 anos) em programa preventivo realizado pela medicina ocupacional. Arq. Bras. Cardiol. [Internet]. 2005 [cited 2021 Mar 19]; 85(1): 20-25. Available from: http://dx.doi.org/10.1590/S0066-782X2005001400005.

6. Cesa CC, Barbiero SM, Pellanda LC. Risco cardiovascular em crianças e adolescentes. Revista da Sociedade de Cardiologia do Estado do Rio Grande do Sul 2010; 20.

7. Avati A, Jung K, Harman S,Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak. 2018;18 (Suppl 4):122.

8. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016; 375(13):1216-1219.

9. Matuchansky C. Deep medicine, artificial intelligence, and the practising clinician. Lancet. 2019; 394(10200), 736.

10. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med. 2020; 26(1):71-76.

11. World Health Organization (WHO). Physical status: the use and interpretation of anthropometry indicators of nutritional status. Geneva: World Health Organization; 1995. (Technical Report Series, 854).

12. Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ. 2007; 85: 660–7.

13. Santos IA, Passos MAZ, Cintra IP, Fisberg M, Ferreti RL, Ganen AP. Pontos de corte de circunferência da cintura de acordo com o estadiamento puberal para identificar sobrepeso em adolescentes. Rev. paul. pediatr. 2019; 37(1): 49-57.

14. Nafiu OO, Burke C, Lee J, Voepel-Lewis T, Malviya S, Tremper KK. Neck circumference as a screening measure for identifying children with high body mass index. *Pediatrics*. 2010; 126(2): e306-e310.

15. Santos HG. Comparação da performance de algoritmos de machine learning para análise preditiva em saúde pública e medicina. 2018. 206f. Tese (Doutorado) - Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo.

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. JMLR. 2011. 12(85):2825−2830.

17. Binieli M. Machine learning: an introduction to mean squared error and regression lines. 2018. Available from: https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/.

18. Dataman D. Explain Your Model with the SHAP Values [Internet]. Medium. Towards Data Science. 2020. Available from: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d.

19. Junge J, Engel C, Vogel M, Naumann S, Löffler M, Thiery J, et al. Neck circumference is similarly predicting for impairment of glucose tolerance as classic anthropometric parameters among healthy and obese children and adolescents. Ann Pediatr Endocrinol Metab 2017; 30(6).

20. Androutsos O, Grammatikaki E, Moschonis G, Roma-Giannikou E, Chrousos GP, Manios Y, et al. Neck circumference: a useful screening tool of cardiovascular risk in children. Pediatric Obesity 2012; 7(3): 187–195.

21. Gomez-Arbelaez D, Camacho PA, Cohen DD, Saavedra-Cortes S, Lopez-Lopez C, Lopez-Jaramillo P. Neck circumference as a predictor of metabolic syndrome, insulin resistance and low-grade systemic inflammation in children: the ACFIES study. BMC Pediatrics 2016; 16(31).

22. De Quadros TMB, Gordia AP, Andaki ACR, Mendes EL, Mota J, Silva LR. Utility of anthropometric indicators to screen for clustered cardiometabolic risk factors in children and adolescents. Ann Pediatr Endocrinol Metab 2019; 32(1):49-55.

23. Montañés EC, Geraud AA, Sardiña NG, Bustos CL. Waist circumference, dyslipidemia and hypertension in prepubertal children. Anales de Pediatria 2007; 67:44-50.

24. Freedman DS, Katzmarzyk PT, Dietz WH, Srinivasan SR, Berenson GS. Relation of body mass index and skinfold thicknesses to cardiovascular disease risk factors in children: the Bogalusa Heart Study. Am. J. Clin. Nutr 2009; 90(1):210-216.

25. Okada T, Sato Y, Yamazaki H, Iwata F, Hara M, Misawa M, et al. Relationship between fat distribution and lipid and apolipoprotein profiles in young teenagers. Pediatrics International 2007; 40(1): 35–40.

26. Kurtoglu S, Hatipoglu N, Mazicioglu MM, Kondolot M. Neck circumference as a novel parameter to determine metabolic risk factors in obese children. European Journal of Clinical Investigation 2011; 42(6): 623–630.

27. Castro-Piñero J, Delgado-Alfonso A, Gracia-Marco L, Gómez-Martínez S, Esteban-Cornejo I, Veiga OL, Marcos A, et. al. Neck circumference and clustered cardiovascular risk factors in children and adolescents: cross-sectional study. BMJ open 2017; 7(9).

28. Ataie-Jafari A, Namazi N, Djalalinia S, Chaghamirzayi P, Abdar ME, Zadehe SS, et al. Neck circumference and its association with cardiometabolic risk factors: a systematic review and meta-analysis. Diabetology & metabolic syndrome 2018.

29. Karatzi K, Moschonis G, Polychronopoulou MC, Chrousos GP, Lionis C, Manios Y. Cut off points of waist circumference and trunk and visceral fat for identifying children with elevated inflammation markers and adipokines: The Healthy Growth Study. Nutrition. 2016; 32(10):1063-1067.