Monitoramento de performance dos membros de academia por estatística descritiva e visualizacão de dados

O que dados como peso e gênero têm a nos dizer sobre a sua performance no treino?

1st Adauta Costa Aragão Freire
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
adautacostaa.freire@alu.ufc.br

2nd Byanca Araújo Pinto
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
byancaa@alu.ufc.br

3rd Carlos Eric Alves Ferreira
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
ferreiraeric@alu.ufc.br

4th Lucas Rodrigues de Azevedo
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
lucas.rodrigues@alu.ufc.br

***Abstract*—Atualmente, com a utilização em massa da internet, percebe-se um grande interesse coletivo pelo uso de dados. Utilizados para o bem ou mal, desenvolvimento de inteligências artificiais, anúncios de lojas ou até mesmo para otimizar tarefas cotidianas. Logo, entende-se que existem diversas abordagens para o uso de dados. Outrossim, em vista do crescimento da informação, nota-se um crescimento na procura de conteúdos que promovem um *lifestyle*, como é popularmente conhecido, mais saudável, o que, consequentemente traz à modernidade uma importância para a vida saudável e bem-estar, sendo um grande portal de entrada para tal estilo de vida a prática de exercícios físicos. Dessarte, utilizando abordagens da estatística descritiva (análise incondicional monovariada, classe-condicional monovariada, incondicional bivariada e incondicional multivariada), buscaremos entender como idade, peso, percentual de gordura e gênero, por exemplo, podem influenciar o corpo e vida das pessoas praticantes de atividades físicas, conforme o dataset utilizado.**

***Index Terms*—análise de dados, visualização de dados, estatística, academia, atividade física, musculação.**

I. Introdução

Com a facilidade e democratização parcial que a internet proporciona, o acesso à ampla disponibilidade de informações se propagou. Sendo assim, é possível visualizar inúmeras páginas de documentos, livros, redes sociais, sites, com apenas uma pesquisa. Mas, para que a utilidade dessas pesquisas seja garantida (ou algo perto disso), é necessário o uso de ferramentas que possam filtrar os dados e identificar padrões relevantes para que assim a busca seja efetiva. Para isso, entretanto, percorre-se um grande caminho até que uma busca seja efetiva (mesmo que as respostas apareçam tão rápido), já que isso envolve um grande campo de pesquisas e conhecimento acumulado ao longo dos anos. Um caminho que engloba diversas áreas de conhecimento, como ciências da computação, estatística, aprendizagem de máquina, inteligência artificial, reconhecimento de padrões, dentre outros.

Dessa forma, é essencial ressaltar que quando algum projeto trabalha com dados, é necessário entender o tipo de dados ideal a ser utilizado e a melhor maneira de usá-los e como os definir. O que chamamos de dados são observações de fenômenos do mundo real, podendo ser preços em uma loja, batimentos cardíacos, pressão arterial, dentre outros [1]. Em diversos projetos que envolvem análise de dados, dificilmente os dados poderão ser usados sem tratamento, já que é comum que alguns valores estejam faltando, sejam redundantes ou não façam sentido, o que causaria um impacto negativo na qualidade da amostra e performance do trabalho. Esse é um dos motivos pelos quais se faz necessário um pré-processamento de dados. O pré-processamento é indispensável em casos em que existem valores não condizentes com o padrão, dados errôneos ou faltosos. Já que a qualidade dos dados é tão importante para que uma análise seja bem sucedida, de maneira alguma poderia existir uma lacuna vazia em uma tabela. Dados com erros ou inexistentes podem contribuir com uma interpretação enviesada ou imprecisa.

Um exemplo prático para isso, é a transformação para resolver *outliers* (valores discrepantes), que normalmente podem receber a definição de dados que possuem informações destoantes do restante da amostra. Podem apresentar erros de medição, eventos incomuns ou variações normais. Em um meio em que se fazem presentes, pode ser um processo desafiador ter que lidar com os *outliers*. Quando uma ou mais amostras são suspeitas de serem *outliers*, o primeiro passo é garantir que os valores são cientificamente válidos [2].

Um exemplo disso pode ser o aferimento dos batimentos cardíacos de uma pessoa, que necessariamente é um valor positivo. No caso, se ele fosse negativo, estaria óbvio que não se trata de um dado verdadeiro, poderia ser o caso de um erro de medição ou registro. Observando o outro lado, se for um valor muito alto, que difere significativamente da maioria dos dados da amostra, poderia se tratar de uma pessoa com a saúde deteriorada, um outlier relevante para o estudo. Portanto, é necessário compreender a precedência dos dados para entender o que fazer com eles. Uma transformação de dados que pode minimizar problemas com os *outliers* é o *spatial sign*, que converte os dados numéricos em uma esfera unitária, multidimensional. É importante centralizar e dimensionar os dados do preditor antes de usar essa transformação.

$$x_{ij}^{*} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{P} x_{ij}^{2}}}.$$

Com o intuito de entender melhor alguns conceitos, serão analisadas rotinas de exercícios, atributos físicos e métricas de condicionamento físico de membros de academia, utilizando ferramentas como analises incondicional monovariada, classe-condicional monovariada, incondicional bivariada e incondicional multivariada.

II. Métodos

A. Descrição do Dataset

O Dataset "Gym Members Exercise" [3] fornece uma visão geral detalhada das rotinas de exercícios, atributos físicos e métricas de condicionamento físico dos membros da academia. Ele contém 973 amostras de dados de academia, incluindo indicadores-chave de desempenho, como frequência cardíaca, calorias queimadas e duração do treino. Cada entrada também inclui dados demográficos e níveis de experiência, permitindo uma análise abrangente dos padrões de condicionamento físico, progressão do atleta e tendências de saúde. Desse modo, foram criados dois subdatasets para representarem 2 classes do Dataset, um para representar o desempenho do gênero masculino e outra do gênero feminino.

As features do dataset são:

1) Idade
2) Gênero
3) Peso
4) Altura
5) Frequência Cardíaca Máxima
6) Frequência Cardíaca Média
7) Frequência Cardíaca em Repouso
8) Duração de Sessão
9) Calorias Queimadas
10) Tipo de Treino
11) Percentual de Gordura
12) Ingestão de Água
13) Frequência de Treino
14) Nível de Experiência
15) Índice de Massa Corporal

B. Análise incondicional monovariada

A aprendizagem estatística refere-se a um conjunto de ferramentas para dar sentido a questões complexas [4]. Mas isso não quer dizer que ela em si seja totalmente complexa, por isso a primeira abordagem estatística a ser utilizada é a análise incondicional monovariada. Inicialmente, sua exploração consiste em observâncias isoladas dos preditores, sem ainda levar em consideração a relação com outras variáveis ou classes. É muito útil para o entendimento das distribuições de cada feature e para identificar padrões que podem inferir análises completas.

Inicialmente, para executar a análise, cria-se um histograma para cada variável, o que permite uma visualização mais clara das distribuições. Outrossim, cálculos de estatística descritiva, como média, desvio padrão e assimetria são feitos. Tais indicadores auxiliam na compreensão de importantes características dos dados obtidos. Outro espectro adicional a ser considerado nesses casos é a identificação de outliers e tipos de distribuição que podem impactar modelos futuros.

Fig. 1. Histograma de Calorias Queimadas.

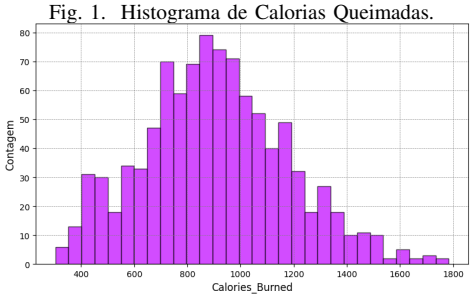


Fig. 2. Histograma de Peso.

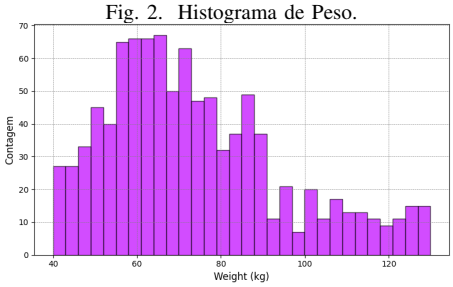
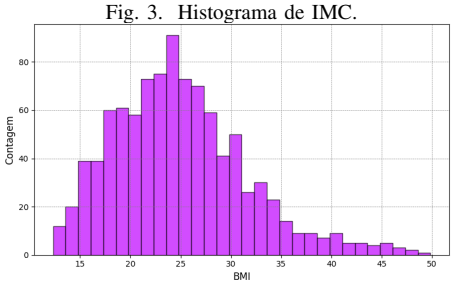


Fig. 3. Histograma de IMC.



Observando os histogramas acima, percebe-se que todos possuem assimetria à direita. No de calorias queimadas, figura

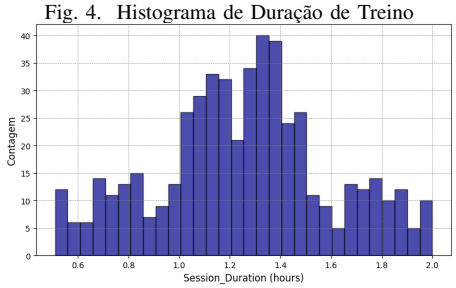
1, percebe-se que a maior parte dos valores está concentrada entre 600kcal e 1200kcal. A maioria das pessoas, conforme a figura 2, possui o peso etre 40kg e 80kg. Já na figura 3, analisando o IMC, a prevalência se encontra entre 20 e 30, o que varia entre a normalidade e o sobrepeso.

C. Análise classe-condicional monovariada

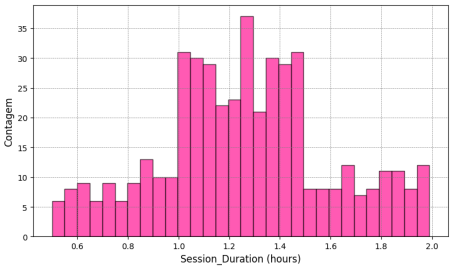
Outra abordagem a ser comentada, é a análise classe-condicional monovariada. Diferentemente da análise incondicional monovariada, aqui as *features* serão analisadas por classes. Como elas foram definidas como ”homem” e ”mulher”, os histogramas azuis serão o dos homens e os rosas, os das mulheres.

Observaremos, então, os histogramas de duração de treino (figuras 4 e 5), peso (figura 7) e índice de massa corporal (figura 6).

Fig. 4. Histograma de Duração de Treino



(a) Masculino



(b) Feminino

De acordo com o histograma acima, percebe-se que os homens tendem a ter uma proporção de sessões mais longas. Todavia, quando se analisa a média do tempo de treino, as mulheres possuem 1,26 hora, enquanto os homens possuem 1,25 hora.

Fig. 5. Histograma de Pesos (Masculino).

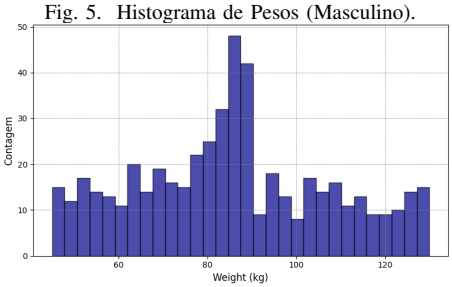
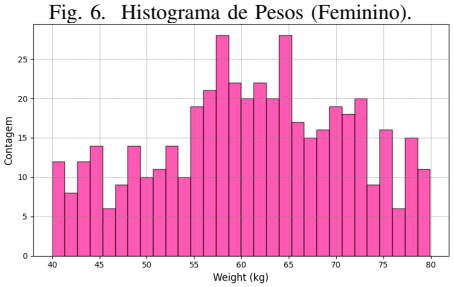


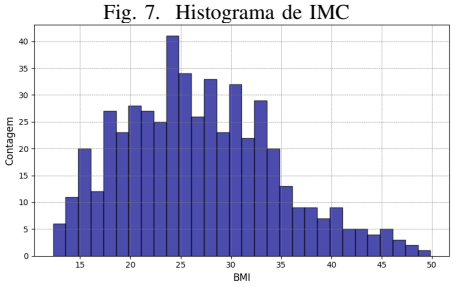
Fig. 6. Histograma de Pesos (Feminino).



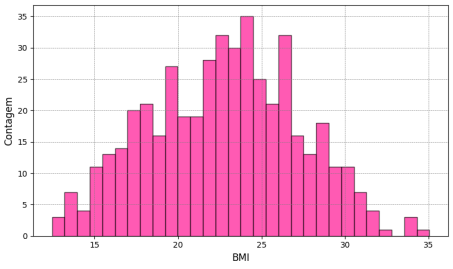
Agora, conforme os histogramas de peso, vistos acima, não é difícil perceber que os homens possuem pesos maiores, com maior parte da distribuição entre 70kg e 100kg, com assimetria à direita. Já as mulheres, possuem uma distribuição mais concentrada entre 50kg e 70kg, também com assimetria à direita.

Conforme analisado anteriormente, o peso das mulheres do dataset tende a ser menor que o dos homens. Sendo assim, é esperado que o IMC (índice de massa corpórea) também seja menor, que o relaciona diretamente com o peso, e mostrado pelos histogramas abaixo.

Fig. 7. Histograma de IMC



(a) Masculino



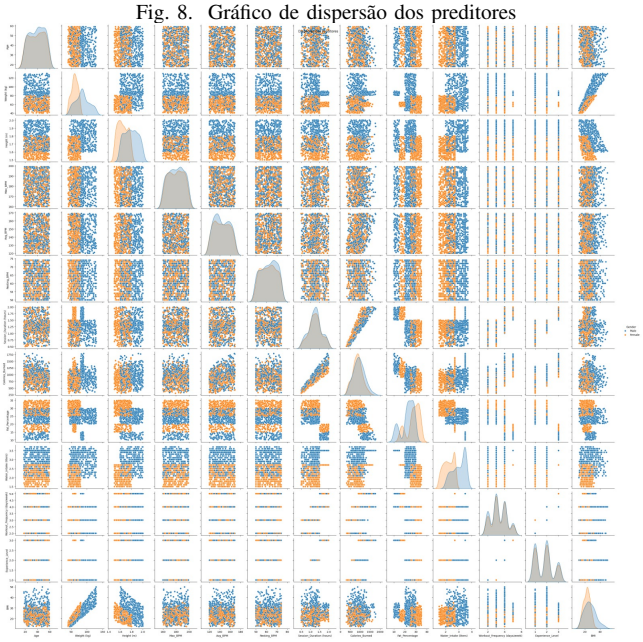
(b) Feminino

D. Análise incondicional bi-variada

A análise incondicional bivariada é a observação da relação entre dois preditores na presença mútua de ambos, sem levar em consideração o impacto de outras variáveis e tem por objetivo identificar associações diretas entre variáveis perceber a influencia de uma feature na outra.

Com base nesse método realizamos a plotagem dos gráficos de todos pontos da relação de pares de preditores, nos quais, definimos a cor laranja para representar a classe ”Mulher” e o azul para a classe ”Homem”.

Fig. 8. Gráfico de dispersão dos preditores



A partir dessa interação, verifica-se a correlação entre as variáveis, permitindo identificar a possibilidade de uma relação positiva ou negativa. Uma correlação perfeita ocorre quando o coeficiente é igual a 1 ou -1, indicando que um preditor influencia totalmente o outro, seja de forma direta ou inversa. Por outro lado, a ausência de correlação é representada pelo coeficiente igual a 0, indicando que uma variável não exerce influência sobre a outra.

Com base nesse entendimento, realizamos a plotagem de todas as correlações, o que nos permitiu visualizar algumas relações relevantes entre os pares de preditores.

Fig. 9. Correlação linear dos preditores

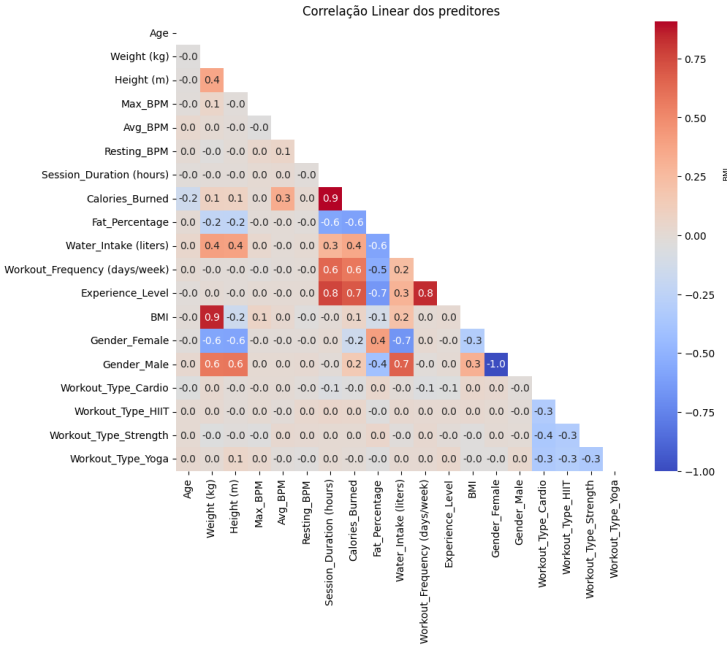
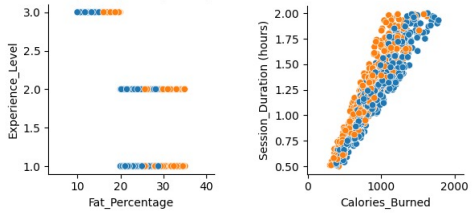
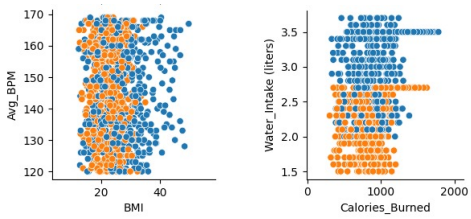


Fig. 10. Gráficos de dispersão com correlaçãoes relevantes



(a) Nível de Experiência x Percentual de Gordura

(b) Duração de Sessão x Calorias Queimadas



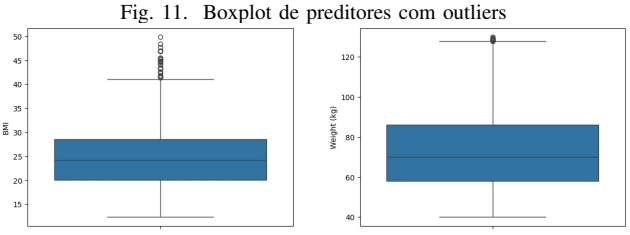
(c) Frequência Cardíaca Média x Índice de Massa Corporal

(d) Ingestão de Água x Calorias Queimadas

Após a análise dos coeficientes de correlação, destacamos alguns gráficos que apresentam relações de maior relevância, refletindo o comportamento já discutido. Para isso, selecionamos quatro exemplos que incluem casos próximos aos valores extremos da correlação, assim como exemplos com coeficientes intermediários, permitindo uma interpretação mais diversificada dos padrões identificados.

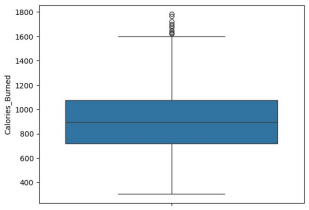
A análise adequada dos dados requer atenção aos outliers presentes nos preditores, já que esses pontos fora do padrão podem impactar os cálculos realizados. Apesar disso, na prática, sua influência nos resultados finais tende a ser limitada. Por esse motivo, optamos por explorar graficamente alguns exemplos de outliers que chamaram a atenção.

Fig. 11. Boxplot de preditores com outliers



(a) Índice de Massa Corporal

(b) Peso



(c) Calorias Queimadas

Analisando as figuras acima, podemos ter uma nova perspectiva sobre os dados, o que possibilita a identificação

de relações entre preditores e potenciais influências entre variáveis.

E. Análise incondicional multivariada

A *Principal Component Analysis* (PCA) é muito utilizada para reduzir a dimensionalidade para simplificar os dados, uma técnica muito utilizada em estatística e aprendizagem de máquina. Além disso, remove correlações entre as variáveis e reduz a complexidade em algoritmos. A PCA procura encontrar, nos preditores, combinações lineares, cuja variância seja a maior possível.

Assim como mencionado anteriormente, a Análise de Componentes Principais cria componentes não correlacionados, sendo uma dessas as principais razões para a sua popularidade. Alguns modelos preditivos optam a ter preditores com baixa ou inexistente correlação com o intuito de melhorar a estabilidade do modelo [1]. Um componente principal j pode ser definidos por:

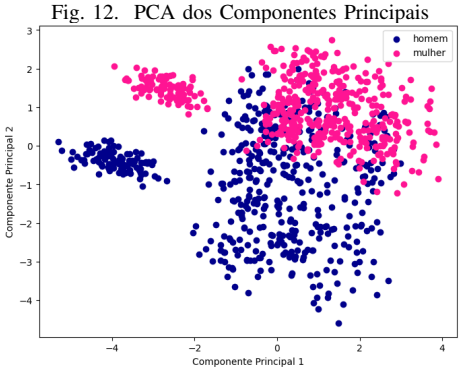
$$\text{PC}_j = a_{j1} \times \text{preditor 1} + a_{j2} \times \text{preditor 2} + \cdots + a_{jp} \times \text{preditor P}$$

Inicialmente, transformamos os dados para que eles fiquem com a média 0 e variância 1. Segue a fórmula abaixo, em que σ é o desvio padrão e μ é a média:

$$\tilde{x}_i = \frac{x_i - \mu}{\sigma}$$

Depois disso, calculamos a matriz de covariância para descobrir os autovalores e autovetores, assim, descobrimos os dois maiores, que são os componentes principais mais importantes. Sendo assim, temos o gráfico abaixo:

Fig. 12. PCA dos Componentes Principais



III. Resultados

A. Resultados da análise incondicional monovariada

Utilizando o dataset, podemos mapear os seguintes dados, utilizando a análise incondicional monovariada:

Estatística é a ciência e arte de tomar decisões baseadas em evidência [5], portanto é necessário analisar minuciosamente antes de falar de resultados. De acordo com os resultados da tabela 1, apresentada abaixo, podemos perceber que os dados apresentados são de adultos, principalmente, já que a média das idades é 38 e o desvio padrão é 12,18. Novamente, o peso, mesmo com o desvio padrão indicando grande variação, mostra-se ser também, em média, de adultos.

TABLE I
Resultados da análise incondicional monovariada

Preditor	μ_d	σ_d	γ_d
Idade	38.68	12.18	-0.077
Peso (kg)	73.85	21.21	0.771
Altura (m)	1.72	0.13	0.338
Frequência cardíaca máxima	179.88	11.55	-0.037
Frequência cardíaca média	143.77	14.34	0.086
Frequência cardíaca em repouso	62.22	7.33	-0.071
Duração da Sessão (horas)	1.25	0.34	0.025
Calorias Queimadas	985.42	272.64	0.277
Percentual de Gordura	24.98	6.25	-0.634
Ingestão de Água (litros)	2.63	0.73	-0.073
Frequência de Treino (dias/semana)	3.32	0.91	0.149
Nível de Experiência	1.81	0.73	0.318
Índice de Massa Corporal (IMC)	24.91	6.66	0.762

Em relação aos dados de peso e altura, mesmo se mostrando valores normais, eles possuem uma leve assimetria positiva, ou seja, alguns valores mais altos. Observando os valores da assimetria, as *features* idade, frequência cardíaca máxima, frequência cardíaca em repouso, duração da sessão e ingestão de água são as que mais se aproximam de uma distribuição simétrica, já que os valores são os que mais se aproximam de zero.

Outrossim, analisando os resultados de calorias queimadas e frequência cardíaca, percebemos uma discrepância entre os dados, o que pode indicar diferentes intensidades nos treinos, o que faz sentido, já que diferentes tipos de exercícios são levadas em consideração nesse dataset, como: treino de força, cardio, yoga e HIIT (*Hight Intensity Interval Training*, Treinamento Intervalado de Alta Intensidade). Além disso, idade, peso, gênero, altura, percentual de gordura, tempo de treino, tudo isso ajuda a definir a quantidade de calorias gastas no exercício e sua frequência cardíaca.

B. Resultados da análise classe-condicional monovariada

TABLE II
Resultados obtidos com a análise classe condicional para o grupo **Feminino**

Preditor	μ_d	σ_d	γ_d
Idade	38.34	12.33	-0.058
Peso (kg)	60.94	10.24	-0.165
Altura (m)	1.64	0.09	0.112
Frequência cardíaca máxima	179.76	11.37	-0.027
Frequência cardíaca média	143.62	14.41	0.057
Frequência cardíaca em repouso	62.11	7.25	-0.005
Duração da Sessão (horas)	1.26	0.34	0.062
Calorias Queimadas	862.25	249.61	0.170
Percentual de Gordura	27.66	5.71	-0.833
Ingestão de Água (litros)	2.21	0.39	-0.185
Frequência de Treino (dias/semana)	3.34	0.90	0.126
Nível de Experiência	1.81	0.74	0.322
Índice de Massa Corporal (IMC)	22.73	4.48	0.001

Conforme a tabela 2 e 3, com os resultados de média, desvio padrão e assimetria dos homens e mulheres, percebe-se,

TABLE III
RESULTADOS OBTIDOS COM A ANÁLISE CLASSE CONDICIONAL PARA O GRUPO **HOMEM**

Preditor	μ_d	σ_d	γ_d
Idade	39.00	12.05	-0.093
Peso (kg)	85.53	21.79	0.152
Altura (m)	1.79	0.12	0.025
Frequência cardíaca máxima	180.00	11.68	-0.048
Frequência cardíaca média	143.90	14.30	0.113
Frequência cardíaca em repouso	62.32	7.40	-0.129
Duração da Sessão (horas)	1.25	0.34	-0.006
Calorias Queimadas	944.46	286.59	0.257
Percentual de Gordura	22.55	5.73	-0.839
Ingestão de Água (litros)	3.01	0.49	-0.458
Frequência de Treino (dias/semana)	3.31	0.93	0.172
Nível de Experiência	1.81	0.74	0.315
Índice de Massa Corporal (IMC)	26.89	7.63	0.461

conforme discutido anteriormente, que as mulheres geralmente possuem o peso e IMC (índice de massa corporal) mais baixo que o dos homens, com média 60,94kg e 22,73kg/m², enquanto o deles é 85,53kg e 26,89kg/m².

Contudo, ao analisar o percentual de gordura, as mulheres se encontram com a média de 27,66%, enquanto os homens possuem, em média, 22,55% de gordura, ambos com assimetria negativa, sendo assim, a maior parte dos valores está concentrada em percentuais mais altos.

Levando em conta esses dados, pode-se dizer, então, que os homens tendem a possuir um percentual maior de massa muscular, pois mesmo o IMC sendo maior, ele apenas calcula a relação entre peso e altura, não definindo a quantidade de gordura e massa magra. Dessa forma, já que eles possuem o IMC mais elevado e menor taxa de gordura, entende-se que os homens possuem mais massa muscular.

C. Resultados da análise incondicional bi-variada

Conforme ilustrado nas figuras 8 e 9, analisamos as relações gráficas e os coeficientes de correlação entre todos os pares de preditores. Entre essas relações, destaca-se a associação entre Duração da Sessão e Calorias Queimadas, que apresenta uma forte correlação positiva de aproximadamente 0.9. Reconhecendo a relevância dessa relação, optamos por gerar um gráfico separado, como mostrado na figura 10(b). A visualização gráfica confirma que essas variáveis possuem uma tendência linear, sendo diretamente proporcionais, com um crescimento conjunto bastante evidente, com valores próximos de um padrão linear.

Em contraste, encontramos uma relação inversa entre o Nível de Experiência e o Percentual de Gordura, com uma correlação negativa de -0.7. Nesse caso, as variáveis mostram um comportamento semelhante ao do exemplo anterior, mas de forma inversamente proporcional, como ilustrado na figura 10(a), onde ambas as variáveis diminuem simultaneamente, mantendo uma linearidade.

Dado o grande número de preditores envolvidos, é possível observar uma diversidade de padrões de relações entre as variáveis, incluindo aquelas que não seguem uma linha reta. Um exemplo claro é a relação entre a Frequência Cardíaca Média e o Índice de Massa Corporal, que apresenta uma correlação de 0, sugerindo que não há influência entre essas variáveis. No gráfico 10(c), observa-se a dispersão dos pontos, sem qualquer tendência linear visível.

Além disso, notamos que a Ingestão de Água tem uma relação positiva com as Calorias Queimadas, com uma correlação de 0.4. Embora essa associação não seja linear, como mostrado no gráfico 10(d), é possível perceber que, em geral, quanto maior a ingestão de água, mais calorias são queimadas, mas sem um aumento proporcional constante.

Essas análises nos permitiram entender as interações entre os preditores, revelando como cada variável pode afetar as outras, seja de forma linear ou não, e identificando padrões de crescimento diretamente ou inversamente proporcionais.

No caso do IMC e do peso, os valores elevados podem indicar casos extremos, como a obesidade. Já nas calorias queimadas, um valor atípico no extremo superior sugere um gasto energético muito maior. Esses outliers devem ser considerados, pois podem influenciar a análise e causar distorções nos modelos ou conclusões.

D. Resultados da análise incondicional multivariada

O gráfico da Análise de Componentes Principais nos mostra os dois principais componentes em um espaço bidimensional. De acordo com a representação dos dados, podemos inferir que existe uma clara separação entre ”homem” e ”mulher”, que estão definidos, respectivamente, nas cores azul e rosa. Isso nos mostra que as características diferem de maneira consistente, o que já era esperado. A separação demonstrada no gráfico pode indicar que variáveis como peso, altura e percentual de gordura podem ter uma clara correlação com o gênero. Ou seja, tais fatores diferenciam significantemente entre as classes.

REFERENCES

[1] Zheng Alice and Casari Amanda. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc, 99

[2] M. Kuhn e K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[3] V. Alakhorasani, ”Gym Members Exercise Dataset,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset. [Accessed: Dec. 3, 2024].

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York, NY: Springer, 2013.

[5] R. M. Heiberger and B. Holland, Statistical Analysis and Data Display: An Intermediate Course with Examples in R. Springer, 2nd ed., 2015.