# Construção de um Modelo de Regressão Logística Ordinal para Classificação do Índice de Massa Corporal

Kalebe Felipe Santana Maia

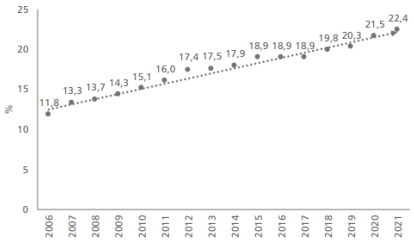
Junho 2024

# Sumário

# 1 Introdução

Obesidade, segundo a Organização Mundial da Saúde(OMS), é definida como o acúmulo em excesso de gordura corporal, sendo uma doença crônica alarmante pelo fato de trazer consigo diversos outros prejuízos para a saúde[1]. Nesse sentido, o desenvolvimento dela decorre de fatores psicossociais, econômicos e culturais mas, apesar de estar vinculada ao aumento de peso, este não indica necessariamente uma consequência da doença. Conquanto estudos antigos sugerem que o sobrepeso, quando combinada com um alto nível de atividade física, está associado a um risco menor na incidência de doenças, outros recentes mostram uma forte relação entre a obesidade e uma tendência ao avanço de um estado não saudável com o passar do tempo[2], apesar de alguns atletas, por exemplo, possuírem um peso elevado.

Atualmente a OMS define o avanço dessa doença não cronológica como epidemia[1] e a classifica como um dos 10 principais problemas de saúde globais, passando a tomar diversas atitudes que objetivam diminuir seu impacto na mortalidade precoce. Essa iniciativa tem um peso maior após analisar os números, visto que cerca de 650 milhões de pessoas, no mundo, são afetadas pela obesidade (1,9 bilhão pelo excesso de peso) e estima-se que 8,8% das mortes no mundo foram causadas pela obesidade [3]. Além disso, no Brasil, a frequência de adultos obesos aumentou consideravelmente, chegando a 22% em 2021 e, entre os estados, o menor número observado foi 17,7%, em Goiânia [4], como mostrado na figura 1.




| CAPITAIS/DF | SEXO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | TOTAL | | | MASCULINO | | | FEMININO | | |
| | % | IC 95% | | % | IC 95% | | % | IC 95% | |
| Aracaju | 25,0 | 20,9 | 29,1 | 24,9 | 18,2 | 31,6 | 25,1 | 20,1 | 30,0 |
| Belém | 25,7 | 21,5 | 29,9 | 22,8 | 16,4 | 29,2 | 28,2 | 22,7 | 33,7 |
| Belo Horizonte | 20,7 | 17,1 | 24,3 | 20,2 | 14,2 | 26,3 | 21,1 | 16,9 | 25,3 |
| Boa Vista | 24,9 | 21,2 | 28,7 | 25,3 | 19,2 | 31,5 | 24,6 | 20,1 | 29,1 |
| Campo Grande | 27,0 | 22,3 | 31,7 | 27,9 | 19,7 | 36,2 | 26,2 | 21,1 | 31,2 |
| Cuiabá | 27,2 | 22,8 | 31,7 | 24,5 | 18,0 | 31,0 | 29,7 | 23,7 | 35,8 |
| Curitiba | 24,5 | 20,5 | 28,5 | 21,9 | 16,0 | 27,8 | 26,7 | 21,3 | 32,1 |
| Florianópolis | 21,9 | 17,8 | 25,9 | 20,6 | 14,8 | 26,5 | 23,0 | 17,4 | 28,6 |
| Fortaleza | 27,7 | 23,2 | 32,1 | 25,2 | 18,1 | 32,4 | 29,8 | 24,2 | 35,3 |
| Goiânia | 17,7 | 14,3 | 21,0 | 19,7 | 14,0 | 25,4 | 15,9 | 12,0 | 19,9 |
| João Pessoa | 22,4 | 18,0 | 26,9 | 21,4 | 14,4 | 28,4 | 23,3 | 17,6 | 29,0 |
| Macapá | 30,4 | 26,0 | 34,9 | 33,4 | 26,6 | 40,3 | 27,5 | 21,8 | 33,3 |
| Maceió | 21,2 | 17,4 | 25,0 | 23,6 | 17,4 | 29,8 | 19,3 | 14,6 | 24,0 |
| Manaus | 27,0 | 22,5 | 31,5 | 26,4 | 18,9 | 33,9 | 27,5 | 22,5 | 32,5 |
| Natal | 21,9 | 18,1 | 25,6 | 22,9 | 16,6 | 29,2 | 21,0 | 16,5 | 25,5 |
| Palmas | 19,0 | 15,5 | 22,4 | 20,9 | 15,3 | 26,5 | 17,3 | 13,0 | 21,5 |
| Porto Alegre | 28,3 | 23,4 | 33,3 | 26,8 | 19,8 | 33,9 | 29,6 | 22,6 | 36,5 |
| Porto Velho | 21,8 | 17,6 | 26,0 | 22,4 | 15,7 | 29,1 | 21,1 | 16,1 | 26,1 |
| Recife | 26,3 | 22,0 | 30,7 | 25,4 | 18,0 | 32,9 | 27,0 | 21,9 | 32,1 |
| Rio Branco | 26,1 | 21,7 | 30,5 | 25,3 | 18,5 | 32,1 | 26,9 | 21,3 | 32,6 |
| Rio de Janeiro | 26,2 | 22,0 | 30,3 | 25,2 | 19,1 | 31,3 | 27,0 | 21,4 | 32,6 |
| Salvador | 25,6 | 21,3 | 29,8 | 24,1 | 17,1 | 31,0 | 26,8 | 21,6 | 32,0 |
| São Luís | 18,5 | 14,8 | 22,3 | 17,3 | 11,3 | 23,3 | 19,5 | 14,8 | 24,2 |
| São Paulo | 24,3 | 20,4 | 28,2 | 25,6 | 19,2 | 31,9 | 23,2 | 18,4 | 28,0 |
| Teresina | 20,8 | 16,9 | 24,8 | 19,8 | 14,1 | 25,5 | 21,7 | 16,2 | 27,2 |
| Vitória | 19,0 | 15,6 | 22,5 | 18,6 | 13,7 | 23,5 | 19,4 | 14,7 | 24,2 |
| Distrito Federal | 21,9 | 17,8 | 26,0 | 16,9 | 11,4 | 22,4 | 26,2 | 20,4 | 32,0 |

*Percentual ponderado para ajustar a distribuição sociodemográfica da amostra Vigitel à distribuição da população adulta de cada cidade projetada para o ano de 2023 (ver Aspectos Metodológicos).
Nota: IC = Intervalo de Confiança de 95%.

Figure 1: Obesidade na população brasileira

Ela também está associada com a diminuição da expectativa e qualidade de vida ao propiciar o surgimento de doenças não transmissíveis perigosas, como diabetes e cânceres[2].

Partindo dessa contextualização, a divisão dos níveis de sobrepeso advém da categorização do cálculo do índice de massa corporal (IMC), pela fórmula

$$\text{IMC} = \frac{\text{peso}}{\text{altura}^2} \tag{1}$$

uma medida internacional utilizada para dizer se uma pessoa está no peso ideal.

Nesse contexto, o objetivo desse trabalho é fazer um modelo preditor utilizando uma regressão logística ordinal(OLR), verificando a causalidade entre alguns fatores e o sobrepeso, para tanto, explicaremos a base teórica deste modelo (juntamente com a interpretação dos coeficientes), a metodologia para seleção de variáveis dependentes através de uma analise exploratória, os resultados do modelo objetivando uma acurácia maior e as conclusões obtidas desse estudo.

Ao final, responder-se-á perguntas do tipo "a variável $X_i$ possui um impacto significativo no avanço entre as categorias de sobrepeso?", bem como perguntas de caráter preditivo, do formato "dado este modelo construído, somos capazes de categorizar o indivíduo com uma confiança alta?". Com base nos dados, espera-se que achemos as variáveis que respondem positivamente a primeira pergunta e, após a construção dos modelos, poderemos verificar a segunda.

# 2 Regressão logística ordinal

Para descrever o modelo, considere uma variável dependente Y que pode ser classificada de J maneiras, sendo que essas categorias possuem uma ordem natural, denotando por $\pi_i$ a probabilidade da categoria i acontecer e sabendo que $\sum_{i=1}^{j} \pi_i = 1$, se existem n observações independentes $y_i$, tem-se então que $y = (y_1, y_2, ..., y_J)$ com $\sum_{j=1} y_j = n$, então a distribuição multinomial, como descrita em [5], é designada por:

$$f(y|n) = \frac{n!}{\prod_{j=1}^{J} y_j} \pi_1^{y_1} \pi_2^{y_2} ... \pi_J^{y_J} \tag{2}$$

Esta distribuição pode ser escrita como $M(n, \pi_1, \pi_2, ..., \pi_J)$ , quando **n** é 2 esta é a clássica distribuição binomial. De um modo geral, a função multinomial 2 não satisfaz os requisitos para ser considerada da família exponencial e, nesse caso, construir o modelo linear generalizado(GLM), como é comum fazer, não é uma solução factível. Para contornar esse fato, pode-se estabelecer uma relação com a distribuição de Poisson.

Com esse objetivo em mente, considere $Y_1, ..., Y_J$ variáveis aleatórias independentes com $Y_i \sim Poisson(\lambda_j)$, então a probabilidade conjunta é:

$$f(y) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \tag{3}$$

Estabelecendo que $n = \sum_{j=1}^{J}$, pode-se dizer então que $n \sim Poisson(\sum_{j=1}^{J} \lambda_j)$.

Com isso, a distribuição de y condicionada a n pode ser escrita como:

$$f(Y|n) = \frac{\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}}{\frac{(\sum_{j=1}^{J} \lambda_j)^n e^{-\sum_{j=1}^{J} \lambda_j}}{n!}} \tag{4}$$

Portanto, podemos nomear $\pi_j = \frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j}$ e, com isso, temos que 2 e 4 são equivalentes. Reduzindo então a distribuição multinomial a uma distribuição conjunta de variáveis Poisson condicionadas a soma delas, podendo assim utilizar o GLM clássico de famílias exponenciais. Ademais, para o modelo multinomial, pode-se mostrar que $\mathbb{E}[Y_j] = n\pi_j$, $Var[Y_j] = n\pi_j(1-\pi_j)$ e $cov(Y_i, Y_j) = -n\pi_i\pi_j$.

Essa contextualização deste espécie de modelo é pressuposto para uma série de modelos onde a variável dependente é classificada em J categorias, como o modelo de chances proporcionais parciais, modelo de razão contínua, modelo esteriótipo e outros descritos em [6]. Para este trabalho, utilizamos o modelo de chances proporcionais(MCP), chamado também de modelo de logito cumulativo, em inglês, *proportional odds model* e *cumulative logit model*, respectivamente. Seu uso é indicado quando a variável dependente era originalmente contínua e depois foi agrupada, o que é exatamente o caso tratado, já que a categorização em níveis de obesidade e sobrepeso provem do agrupamento de intervalos de IMC, uma variável contínua.

Para utilizá-lo, existem alguns pressupostos. Em primeiro lugar, avalia-se o tipo da variável resposta, ou seja, um pressuposto teórico de que ela seja categórica ordinal, o que é verificado antes da escolha do modelo e, de fato, possui essa característica. O segundo pressuposto é a independência das observações, ou seja, esse pressuposto é avaliado na prática e, nesse caso, o dataset é construído de forma que cada entrada seja uma pessoa diferente. Por fim, pede-se ainda a ausência de multicolinearidade.

Esse último pressuposto é facilmente rompido e, caso isso aconteça, o certo seria fazer uma regressão multinomial. O teste para este requisito compara se os coeficientes são muito diferentes entre as categorias ou não, ao relaxar a condição de chances proporcionais. Por fim, nesse modelo, diferentemente de alguns outros citados, acredita-se que o impacto de uma variável independente é o mesmo entre as classes, ou seja, a idade impacta no nível de obesidade 1 para 2 na mesma quantidade que impacta no peso normal para o sobrepeso 1, isto é, o impacto em categorias adjacentes é o mesmo. Com isso, já sabemos grande parte dos fundamentos dele, vamos então ver como tratar os resultados obtidos.

## 2.1 Odds ratio

A saída desse modelo pode ser descrita como:

$$ln(\frac{P(Y \leq j)}{P(Y > j))}) = \zeta - \eta_1 X_1 - \eta_2 X_2 - ... - \eta_n X_n \tag{5}$$

Sendo n o número de variáveis independentes, j cada categoria possível até $J-1$, caso contrário, $P(Y > j) = 0$ e acontece uma divisão por 0, o que não seria interessante. Como na regressão logística clássica, o lado esquerdo é o log-odds de uma probabilidade, mas agora da probabilidade cumulativa.

Podemos interpretar $\zeta$ como um intercepto que representa o log-odds de $Y \leq j$ quando todos preditores são 0, ou seja, $P(Y \leq j)$ é o logito inverso de zeta, possuindo um intercepto para cada categoria entre as possíveis classificações (1 até $J-1$). Cada $-\eta_i$ para $i \in \{1, ..., n\}$ é o log do odds ratio (razão de chances) comparando as chances de $Y \leq j$ entre indivíduos que diferem em uma unidade em $X_i$. Com isso, escrevemos $e^{-\eta_i}$ como o odds ratio comparando as chances de $Y \leq j$ entre $X_i$ que diferem em uma unidade e $e^{\eta_i}$ o odds ratio comparando as chances de $Y > j$ para entre $X_i$ que diferem em uma unidade. Concluindo que toda interpretação no modelo tera isso como base.

Apenas sintetizando essa última parte, não há grande diferença entre a interpretação da razão de chances entre um modelo de regressão logística e um modelo de regressão logística ordinal com o pressuposto de chances proporcionais. Portanto, se exponenciarmos os etas e verificarmos, por exemplo, um valor de 1.5 atrelado a uma variável independente qualquer, inferimos então que a cada variação unitária dessa variável, mantendo todas as outras constantes, há uma chance de 50% de uma pessoa em peso normal passar para o sobrepeso 1, de uma pessoa em sobrepeso 1 passar para sobrepeso 2 e assim por diante, afinal, isso quer dizer chances proporcionais, sendo a única diferença notada entre os interceptos, que são feitos em comparação com a classe anterior e existe um para cada categoria.

## 2.2 Qualidade do ajuste

Na avaliação do modelo, pode-se utilizar o critério de informação de Akaike (AIC), uma métrica que leva em consideração a simplicidade do modelo, onde valores menores são melhores. Essa informação pode ser uma ferramenta útil na escolha do modelo, sendo calculada por:

$$AIC = 2K - 2log(L(\hat{\theta})) \tag{6}$$

onde K o número de parâmetros do modelo, n é o tamanho da amostra em questão e L a log-verossimilhança avaliada no $\hat{\theta}$.

Além disso, temos também, por ser um modelo preditivo, o instrumento da matriz de confusão. Ela é descrita como:

| | | Valor Predito | |
|---|---|---|---|
| | | **Sim** | **Não** |
| **Real** | **Sim** | Verdadeiro Positivo (TP) | Falso Negativo (FN) |
| | **Não** | Falso Positivo (FP) | Verdadeiro Negativo (TN) |

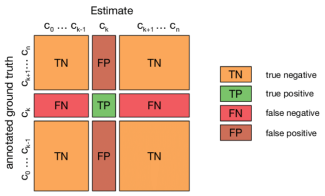



Figure 2: Matrizes de confusão

A figura 2 basicamente mostra a relação entre os valores preditos e os verdadeiros valores, a primeira tabela equivale a clássica regressão logística, aonde as possíveis saídas pertencem a 2 classes, enquanto a segunda, o nosso caso, avalia quando há mais de 2 categorias possíveis de serem preditas. Existem 4 possibilidades (ao avaliar determinada classe i, com $i, j \in \{1, 2, ..., J\}$), sendo elas:

1. verdadeiro positivo(TP), os valores são da classe i e foram classificados como pertecentes a classe i,
2. falso positivo(FP), os valores são da classe j mas foram classificados como da classe i,
3. falso negativo(FN), os valores são da classe i mas foram classificados como de classe j,
4. verdadeiro negativo(TP), os valores não são da classe i e não foram classificados como da classe i.

Como descrito em [7], existem alguns valores de interesse ao olhar a matriz de confusão. Primeiro, temos a acurácia, uma medida que fala sobre a probabilidade do modelo acertar previsões, seu cálculo é dado por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Em segundo, a sensibilidade, um indicador cujo objetivo é apresentar o quão bom o modelo é na previsão de sucessos, dado pela fórmula:

$$\text{Sensibilidade} = \frac{TP}{TP + FN} \tag{8}$$

Em terceiro, a especificidade, ao contrário da sensibilidade, ele indica o quão bom o modela é na previsão de fracassos, dado por:

$$\text{Especificidade} = \frac{FP}{TN + FP} \tag{9}$$

Para esse modelo, e em um contexto onde a predição envolve doenças, a escolha de priorizar a precisão deve ser interessante, pois evita casos em que o

diagnóstico é dito como negativo e na verdade a pessoa porta a doença, o que pode trazer complicações futuras, visto que grande parte das doenças tratadas em seu principio tendem a ter um menor impacto na vida da pessoa.

# 3 Modelagem

Para construir nosso modelo, devemos primeiro descobrir quais variáveis colocar. Para tanto, temos um conjunto de dados[8] com 14 colunas (variáveis independentes) e 2111 linhas (amostras), sendo a nossa variável dependente a categoria em que uma dada pessoa se encontra em relação ao peso. Temos a seguinte tabela dos preditores:

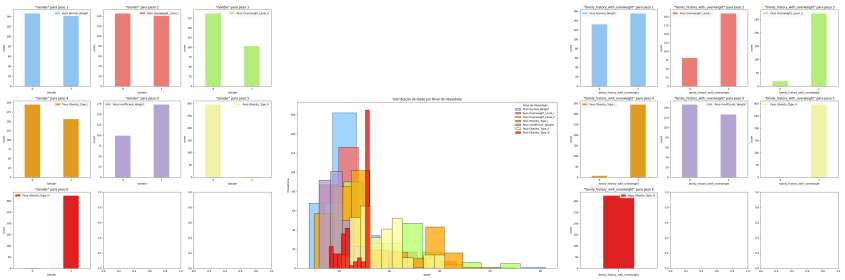
A variável dependente é a categoria do peso baseado no IMC e possui os seguintes valores: Insufficient_Weight (transformado em 0), Normal_Weight (transformado em 1), Overweight_Level_I (transformado em 2), Overweight_Level_II (transformado em 3), Obesity_Type_I (transformado em 4), Obesity_Type_II (transformado em 5), e Obesity_Type_III (transformado em 6)

## 3.1 Análise Exploratória

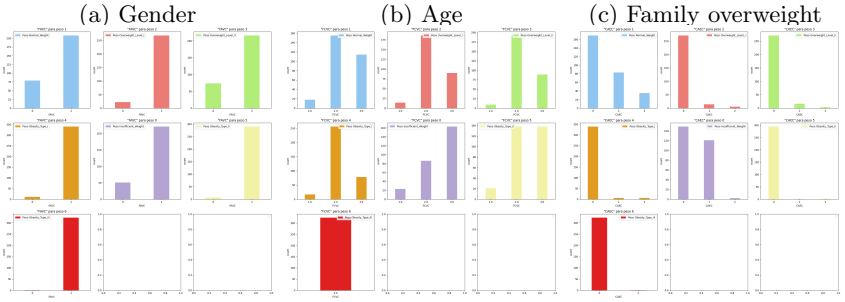
Em primeiro lugar, como só existe uma única variável contínua, não podemos visualizar scatter plots, algo um pouco ruim, mas vamos trabalhar com as frequências de incidência de cada classe da obesidade em relação a alguns parâmetros, vejamos os seguintes plots (esses possuem valores de interesse), gerados através do código disponível github.com/kalebemaiaa/Modelagem-Estatsica:

| Variável | Descrição | Valores | Tipo |
|---|---|---|---|
| Gender | Gênero de cada pessoa: | homens (Male) = 0<br>mulheres (Female) = 1 | Binária |
| Age | Idade de cada pessoa | float | Contínua |
| family_history_with_overweight | Histórico de sobrepeso na família: | sim = 1<br>não = 0 | Binária |
| FAVC | Frequência de consumo de comidas hipercalóricas: | frequente = 1<br>caso contrário = 0 | Binária |
| FCVC | Frequência de consumo de vegetais: | nunca = 0<br>algumas vezes = 1<br>sempre = 2 | Categórica |
| NCP | Número de refeições principais: | entre 1 e 2 refeições = 0<br>3 refeições = 1<br>mais que 3 = 2 | Categórica |
| CAEC | Consumo de comidas entre as refeições: | não consome = 0<br>algumas vezes = 1<br>frequentemente = 2<br>sempre = 3 | Categórica |
| SMOKE | Pessoa fumante: | sim = 1<br>não = 0 | Binária |
| CH20 | Consumo de água diário: | menos que 1 litro = 0<br>entre 1 e 2 litros = 1<br>mais que 2 litros = 2 | Categórica |
| CALC | Consumo de álcool: | não bebe = 0<br>algumas vezes = 1<br>frequentemente = 2<br>sempre = 3 | Categórica |
| SCC | Monitora a quantidade de calorias ingeridas: | sim = 1<br>não = 0 | Binária |
| FAF | Frequência semanal da prática de atividades físicas: | não pratica = 0<br>entre 1 e 2 dias = 1<br>entre 2 e 4 dias = 2<br>mais que 4 dias = 3 | Categórica |
| TUE | Tempo de uso de tecnologias: | entre 0 a 2 horas = 0<br>entre 3 a 5 horas = 1<br>mais que 5 horas = 2 | Categórica |
| MTRANS | Método de locomoção principal | andando = 0<br>transporte público = 1<br>bicicleta = 2<br>moto = 3<br>carro = 4 | Categórica |

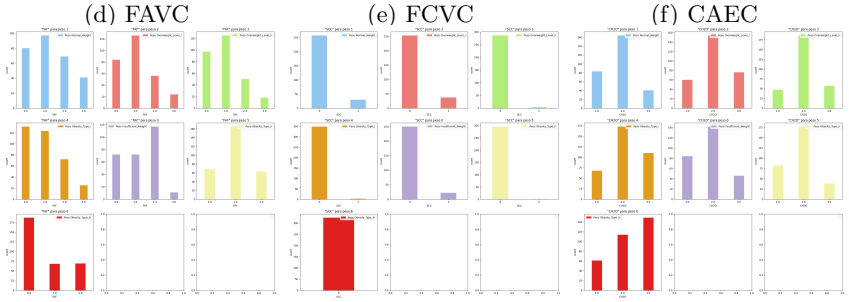
Table 1: Descrição das variáveis



(a) Gender (b) Age (c) Family overweight



(d) FAVC (e) FCVC (f) CAEC



(g) FAF (h) SCC (i) CH2O

Figure 3: Plots frequência de interesse

Primeiro, em gênero, os valores se mostram quase que perfeitamente balanceados para pesos normais até que, em um dado momento, há quase que exclusivamente pessoas do gênero 1 (female) em situações de obesidade em estágio 3 e pessoas do gênero 0 (male) em obesidade do tipo 2, de resto, não há uma predominância.

Outro fator dominante é, sem duvida alguma, o histórico de sobrepeso na família. Esses plots mostram que, a partir do estágio 1 de sobrepeso, há uma prevalência de histórico de sobrepeso na família e, pior ainda, a partir do sobrepeso de nível 2, quase que 90% das pessoas apresentam essa tendência, chegando em 100% quando as pessoas apresentam obesidade de tipo 3. Esses fatos mostram que essa é uma variável em que possuímos interesse e pode vir a revelar que há uma perpetuação do sobrepeso na família.

Notemos ainda que a frequência de consumo de alimentos hipercalóricos possui valor majoritariamente 1 em quase todas as pessoas, mas nas pessoas em estágio de obesidade de nível 1, nível 2 e nível 3, esse valor é, em comparação

com os demais, mais alto.

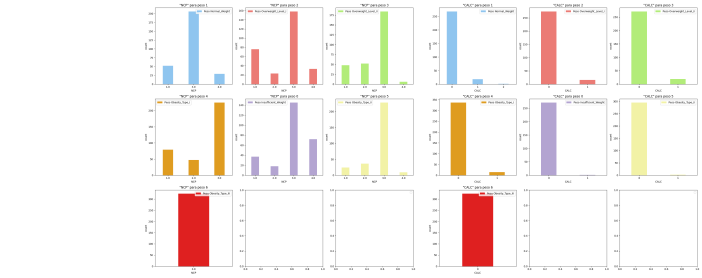
Com relação ao consumo de vegetais, há um questionamento a ser feito, visto que quem está em sobrepeso possui uma tendência a comer frequentemente vegetais. Nesse sentido, a duvida é "pessoas em sobrepeso tendem a comer mais de tudo?", o que é uma questão plausível, ou simplesmente "não há relação entre o consumo de vegetais e uma saciedade maior acompanhada de um menor ganho calórico?", fazendo com que mesmo o consumo maior de vegetais ainda gere um ganho de peso relevante.

Com relação ao consumo de comidas entre as refeições, vemos que há algo contraditório ao que é esperado, já que grande parte de quem está em estágios avançados de sobrepeso relatam não comerem entre as refeições. Para entender isso, podemos combinar essa informação com o plot de SCC, ou seja, o monitoramento da quantidade de calorias consumidas. Um fato constatado pelos nutricionistas [9] é que há uma sub notificação das pessoas que fazem dieta, isto é, muitas vezes, por não monitorarem a quantidade de calorias consumidas ao longo do dia, esquecem de reportar o consumo entre as refeições principais e, por isso, não notam que consomem mais do que o que é indicado.

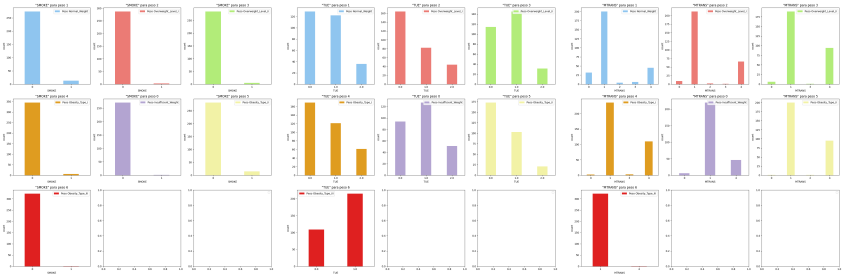
Nesse contexto, notamos uma tendência das pessoas reportarem determinados valores nesses quesitos, no caso, dizerem que não consomem nada entre as refeições e, apesar de parecer contraintuitivo, a verdade pode ser outra. Mas, pensando dessa forma, qualquer resposta obtida pode ser fruto de uma mentira inconsciente e, se for assim, quaisquer considerações feitas não irão valer, para evitar isso, vamos apenas dizer que a interferência é negativa, ou seja, pessoas em sobrepeso tendem a consumir menos comida entre as refeições principais.

Ademais, a frequência de prática de atividades físicas só possui uma diferença visível a partir do estágio 1 de obesidade, onde a frequência de mais de 4 dias na semana de práticas esportivas cai bastante, em comparação aos demais, chegando a sumir no estágio 2 e 3 e, além disso, no estágio 3, verifica-se que a maior parte não pratica estas atividades em nenhum dia da semana.

Por fim, um comentário a ser feito é em relação ao consumo de água, já que grande parte das pessoas com obesidade em estágio 3 consomem mais de 2 litros de água por dia, o que é um hábito comprovadamente saudável. Por outro lado, estudos sugerem que há uma ligação entre indivíduos obesos não estarem hidratados [10] com maior facilidade, com isso, há a necessidade de aumentar o consumo de água. Estes fatos levantam o questionamento da relação causal, afinal, a pessoa estar obesa implica em um aumento do consumo de água ou um maior consumo de água implica em uma maior tendência ao aumento do IMC? A priori, parece intuitivo que é a primeira opção e, por isso, baseando-me também na relação da obesidade com o desenvolvimento de diabetes de tipo 2 e, está última, com a maior necessidade de consumo de água, essa variável não sera levada em consideração no modelo construído.



(a) NCP (b) CALC



(c) SMOKE (d) TUE (e) MTRANS

Figure 4: Plots de frequência sem informações relevantes

Com relação a figura 4, diferentemente da anterior, esses plots não mostram quase nenhuma relação com a prevalência e avanço do sobrepeso, já que os valores são quase idênticos para todas as classes. Sendo eles: o indicador de fumo (quase idêntico para todos), o meio de transporte usual, o tempo de uso de apetrechos tecnológicos (não há variância significativa entre as classes), número de refeições principais (quase todas as classes consomem mais de 3 refeições) e consumo de álcool (em todas as classes, grande parte não consome). Lembrando que as classes aqui ditas são referentes as classes da variável dependente.

Além desses plots, vamos analisar a covariância entre as variáveis independentes para verificar se devemos adicionar alguma interação no nosso modelo.

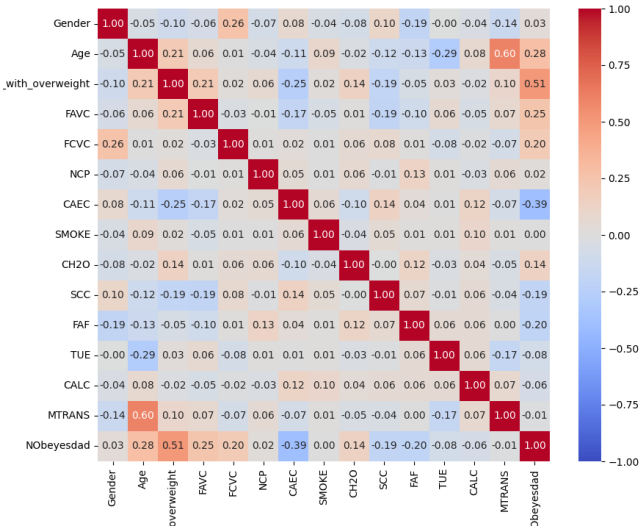



Figure 5: Matriz de covariância

Vendo a figura 5, percebemos que, com relação as variáveis independentes, há uma possível correlação entre :

1. a idade e o método de transporte escolhido
2. o gênero da pessoa e o consumo de vegetais
3. a idade e o tempo de uso de aparelhos tecnológicos
4. a frequência da prática de esportes físicos e o gênero
5. o histórico família de sobrepeso e o consumo de alimentos calóricos
6. a frequência de consumo de alimentos hipercalóricos com o consumo de comidas entre as refeições

No geral, a primeira interação pode ser ignorada já que, o conhecimento prévio de estudos, mostra que a relação de causalidade é: a idade determina o meio de transporte escolhido. Sendo assim, como não há uma implicação entre o meio de transporte e o sobrepeso, pode-se dizer que não há porque colocar essa interação no modelo. Por esse motivo também, não há porque colocar a interação 3. Uma possível explicação mais elaborada de porque não escolher estas interações envolveria construir o DAG(Grafo Acíclico Dirigido) de dependências causais e observar que elas não impactam no resultado que queremos regredir.

Pensando também em implicações causais, as outras variáveis, a priori, não aparentam serem fatores de confusão. Ademais, esse heatmmap, em um contexto geral, deixa evidente a correlação entre algumas variáveis anteriormente

analisadas e nossa variável dependente. Sendo assim, podemos adicionar algumas destas correlações em alguns de nossos modelos.

# 4 Resultados

## 4.1 Modelos escolhidos

Para o primeiro modelo, chamaremos de M1, vamos escrever:

$$\begin{aligned} Y_i = \zeta &- \eta_1 \text{FHWO} - \eta_2 \text{FAVC} - \eta_3 \text{Age} - \eta_4 \text{FCVC} - \eta_5 \text{CAEC} \\ &- \eta_6 \text{Gender} - \eta_7 \text{FAF} \end{aligned} \tag{10}$$

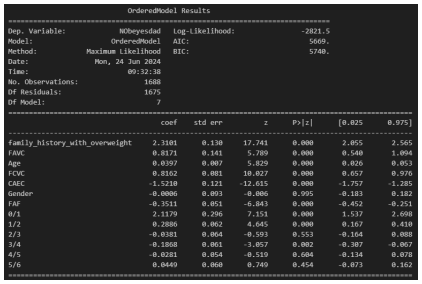
Para o segundo modelo, chamaremos de M2, vamos incluir algumas interações, no geral, todas que não foram excluídas da lista, e escreveremos ele como:

$$\begin{aligned} Y_i = \zeta &- \eta_1 \text{FHWO} - \eta_2 \text{FAVC} - \eta_3 \text{Age} - \eta_4 \text{FCVC} - \eta_5 \text{CAEC} \\ &- \eta_6 \text{Gender} - \eta_7 \text{FAF} - \eta_8 \text{Gender} * \text{FCVC} \\ &- \eta_9 \text{Gender} * \text{FAF} - \eta_{10} \text{FHWO} * \text{FCVC} - \eta_{11} \text{FCVC} * \text{CAEC} \end{aligned} \tag{11}$$

Aqui houve um abuso de notação para simplificar a relação modelada, mas esses etas são o log das razões de chances (no geral, essa função é do formato descrito em 5). Ademais, poderíamos fazer modelos intermediários, adicionando somente algumas interações, mas por limitação de espaço e outros fatores, esses 2 devem servir para um bom começo. Antes dos resultados em si, espera-se que, em comparação com 10, o modelo 11 possua um valor de $\eta_6$ menor, já que seu impacto será transferido para $\eta_{10}$ e $\eta_9$. Pelo mesmo motivo, $\eta_4$ deve ser menor também.

## 4.2 Resultados numéricos

A implementação foi feita através do módulo statsmodels da linguagem python. Obtivemos os seguinte retorno dos modelos:

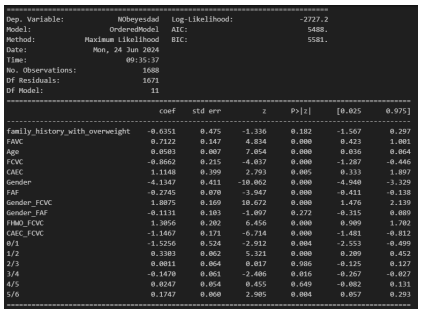


OrderedModel Results

| | | | |
|---|---|---|---|
| Dep. Variable: | NObeyesdad | Log-Likelihood: | -2821.5 |
| Model: | OrderedModel | AIC: | 5669. |
| Method: | Maximum Likelihood | BIC: | 5740. |
| Date: | Mon, 24 Jun 2024 | | |
| Time: | 09:32:38 | | |
| No. Observations: | 1688 | | |
| Df Residuals: | 1675 | | |
| Df Model: | 7 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| family_history_with_overweight | 2.3101 | 0.130 | 17.741 | 0.000 | 2.055 | 2.565 |
| FAVC | 0.8171 | 0.141 | 5.789 | 0.000 | 0.540 | 1.094 |
| Age | 0.0397 | 0.007 | 5.829 | 0.000 | 0.026 | 0.053 |
| FCVC | 0.8162 | 0.081 | 10.027 | 0.000 | 0.657 | 0.976 |
| CAEC | -1.5210 | 0.121 | -12.615 | 0.000 | -1.757 | -1.285 |
| Gender | -0.0006 | 0.093 | -0.006 | 0.995 | -0.183 | 0.182 |
| FAF | -0.3511 | 0.053 | -6.843 | 0.000 | -0.452 | -0.251 |
| 0/1 | 2.1179 | 0.296 | 7.151 | 0.000 | 1.537 | 2.698 |
| 1/2 | 0.2886 | 0.062 | 4.645 | 0.000 | 0.167 | 0.410 |
| 2/3 | -0.0381 | 0.064 | -0.593 | 0.553 | -0.164 | 0.088 |
| 3/4 | -0.1868 | 0.061 | -3.057 | 0.002 | -0.307 | -0.067 |
| 4/5 | -0.0281 | 0.054 | -0.519 | 0.604 | -0.134 | 0.078 |
| 5/6 | 0.0449 | 0.060 | 0.749 | 0.454 | -0.073 | 0.162 |

(a) Resultados modelo 1



| | | | |
|---|---|---|---|
| Dep. Variable: | NObeyesdad | Log-Likelihood: | -2727.2 |
| Model: | OrderedModel | AIC: | 5488. |
| Method: | Maximum Likelihood | BIC: | 5581. |
| Date: | Mon, 24 Jun 2024 | | |
| Time: | 09:35:37 | | |
| No. Observations: | 1688 | | |
| Df Residuals: | 1671 | | |
| Df Model: | 11 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| family_history_with_overweight | -0.6351 | 0.475 | -1.336 | 0.182 | -1.567 | 0.297 |
| FAVC | 0.7122 | 0.147 | 4.834 | 0.000 | 0.423 | 1.001 |
| Age | 0.0503 | 0.007 | 7.054 | 0.000 | 0.036 | 0.064 |
| FCVC | -0.8062 | 0.215 | -4.017 | 0.000 | -1.287 | -0.446 |
| CAEC | 1.1148 | 0.399 | 2.793 | 0.005 | 0.333 | 1.897 |
| Gender | -4.1347 | 0.411 | -10.062 | 0.000 | -4.940 | -3.329 |
| FAF | -0.2245 | 0.070 | -3.947 | 0.000 | -0.411 | -0.138 |
| Gender_FCVC | 1.0075 | 0.169 | 10.672 | 0.000 | 1.476 | 2.139 |
| Gender_FAF | -0.1131 | 0.103 | -1.097 | 0.272 | -0.315 | 0.089 |
| FHWO_FCVC | 1.3056 | 0.202 | 6.456 | 0.000 | 0.909 | 1.702 |
| CAEC_FCVC | -1.1467 | 0.171 | -6.714 | 0.000 | -1.481 | -0.812 |
| 0/1 | -1.5256 | 0.524 | -2.912 | 0.004 | -2.553 | -0.499 |
| 1/2 | 0.3383 | 0.062 | 5.321 | 0.000 | 0.209 | 0.452 |
| 2/3 | 0.0011 | 0.064 | 0.017 | 0.986 | -0.125 | 0.127 |
| 3/4 | -0.1470 | 0.061 | -2.406 | 0.016 | -0.267 | -0.027 |
| 4/5 | 0.0247 | 0.054 | 0.455 | 0.649 | -0.082 | 0.131 |
| 5/6 | 0.1747 | 0.060 | 2.905 | 0.004 | 0.057 | 0.293 |

(b) Resultados modelo 2

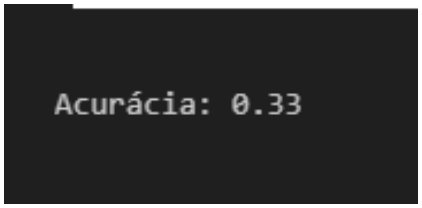
Figure 6: Resultados dos Modelos M1 e M2

Para este cálculo, utilizamos a função *OrderedModel* com a função de ligação **logit**, além disso, para o fit do modelo, foi utilizado o parâmetro *newton* que se refere ao método de Newton-Raphson.

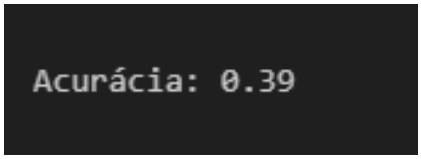
## 4.3 Interpretando resultados

Antes de mais nada, vemos que o AIC do segundo modelo foi menor e, como citado, quanto menor for esse valor, preferível é o modelo (olhando somente esse parâmetro). Além desse valor, também obtivemos o BIC e o Log-likelihood, que não serão objeto de discussão, mas o modelo 2 funciona melhor que o 1, baseado nesses valores também, apesar de o Log-likelihood não estar em um número bom para nenhum dos dois (isso indica um mal ajuste do modelo ao conjunto de dados).

Seguindo em diante, cabe explicar que o dataset foi particionado para teste, sendo 80% usado para descobrir os coeficientes e 20% usado para verificar a acurácia do modelo, com isso, obtivemos os seguintes resultados:



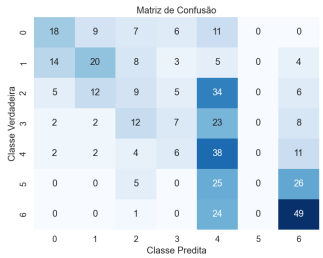

(a) Acurácia Modelo 1



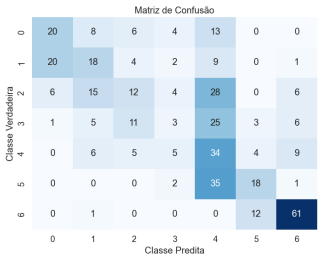

(b) Acurácia Modelo 2

Figure 7: Comparação das Acurácias dos Modelos M1 e M2

Esses resultados não eram os planejados tanto para o primeiro modelo, quanto para o segundo. Ao contrário do que foi obtido, era esperado que a capacidade de predição fosse relativamente mais alta e, mais que isso, a diferença entre os dois fosse mais significativa. Vamos então verificar as matrizes de confusão:




(a) Matriz de Confusão - Modelo M1




(b) Matriz de Confusão - Modelo M2

Figure 8: Matrizes de Confusão dos Modelos M1 e M2

Nesse caso, apesar das acurácias extremamente baixas, vemos que o modelo 2 apresenta um número alto de verdadeiros positivos de pessoas em estágio 3 de obesidade, o que é algo bom, no sentido de antecipar esse vento, já que este é um estágio grave e avançado da doença. Um problema facilmente inidentificável é que ambos erraram bastante em se tratando da predição da classe 5 e da classe 4, isto é, a obesidade em nível 1 foi predita muitas vezes, quando não era verdade, enquanto a de nível 2 foi predita poucas vezes, quando na realidade deveria ser mais vezes. Apesar dos problemas evidentes, quase não há casos em que a categoria predita fosse menor do que a verdadeira, com isso, se por um lado há uma porcentagem gritante de erros na predição, por outro, ao menos no modelo 2, as pessoas não são preditas em estágio de peso normal ou sobrepeso inicial quando na verdade estão em estado de obesidade.

Apesar disso, os resultados ainda se mostram bem ruins. Vamos então verificar as razões de chance e ver se alguns resultados convém com a análise exploratória e algum efeito se mostrou impactante de verdade.

| | Odd ratio | Lower bound | Upper bound |
|---|---|---|---|
| family_history_with_overweight | 10.075446 | 7.805976 | 13.004730 |
| FAVC | 2.263872 | 1.716767 | 2.985331 |
| Age | 1.040474 | 1.026684 | 1.054448 |
| FCVC | 2.261933 | 1.928356 | 2.653214 |
| CAEC | 0.218497 | 0.172509 | 0.276743 |
| Gender | 0.999430 | 0.832795 | 1.199407 |
| FAF | 0.703883 | 0.636534 | 0.778357 |
| 0/1 | 8.313440 | 4.652442 | 14.855270 |
| 1/2 | 1.334533 | 1.181531 | 1.507348 |
| 2/3 | 0.962591 | 0.848664 | 1.091813 |
| 3/4 | 0.829608 | 0.735971 | 0.935159 |
| 4/5 | 0.972296 | 0.874383 | 1.081173 |
| 5/6 | 1.045932 | 0.929927 | 1.176409 |

(a) Odds Ratio M1

| | Odd ratio | Lower bound | Upper bound |
|---|---|---|---|
| family_history_with_overweight | 0.529879 | 0.208683 | 1.345444 |
| FAVC | 2.038376 | 1.527144 | 2.720750 |
| Age | 1.051547 | 1.036963 | 1.066335 |
| FCVC | 0.420537 | 0.276167 | 0.640379 |
| CAEC | 3.049030 | 1.394590 | 6.666177 |
| Gender | 0.016007 | 0.007154 | 0.035816 |
| FAF | 0.759934 | 0.663083 | 0.870930 |
| Gender_FCVC | 6.094926 | 4.373247 | 8.494403 |
| Gender_FAF | 0.893044 | 0.729681 | 1.092982 |
| FHWO_FCVC | 3.690059 | 2.482571 | 5.484852 |
| CAEC_FCVC | 0.317691 | 0.227311 | 0.444005 |
| 0/1 | 0.217480 | 0.077875 | 0.607353 |
| 1/2 | 1.391426 | 1.232021 | 1.571456 |
| 2/3 | 1.001118 | 0.882776 | 1.135324 |
| 3/4 | 0.863264 | 0.765804 | 0.973126 |
| 4/5 | 1.025010 | 0.921640 | 1.139975 |
| 5/6 | 1.190947 | 1.058514 | 1.339949 |

(b) Odds Ratio M2

Figure 9: Comparação dos Odds Ratio entre Modelos

Pela tabela, intervalos onde o número 1 esteja entre o limite inferior e o limite superior não podem nos dar uma inferência valida e nada podemos afirmar sobre eles, seja positivamente ou negativamente.

Em primeira instância, é observável que a maior diferença entre os dois modelos está no variável que indica o histórico de sobrepeso familiar, visto que há uma redução, onde, no primeiro modelo, se a indicadora é positiva, há 10 vezes mais chances de uma pessoa transitar para outra categoria mais avançada, mantendo as outras variáveis constantes, já no segundo, desse valor não inferimos nada, apesar de, como explicado e previsto, esse número ter diminuído por absorvido pela interação com a frequência de consumo de vegetais. Seguindo essa linha de analise, vê-se que de fato a interação entre o histórico de sobrepeso familiar com o consumo de vegetais em grande frequência possui um efeito positivo, aumentando em 260% a chance de uma pessoa transitar para a próxima categoria de peso.

Um impacto que quase não variou foi o da idade, sendo que, se compararmos duas pessoas onde todos atributos são constantes e apenas a idade varia, podemos dizer que a probabilidade de quem possui uma idade mais avançada pertencer a próxima classe de sobrepeso aumenta aproximadamente 4%, no primeiro modelo, e 5%, no segundo, para cada ano a mais que uma pessoa possui. Além desse, a frequência de consumo de alimentos hipercalóricos seguiu essa mesma tendência, mantendo as chances em torno de 2 vezes da pessoa transitar para a próxima categoria em ambos casos, sendo pouco maior no primeiro.

Como esperado, e dito anteriormente, o coeficiente da frequência de consumo de vegetais diminui bastante do modelo 1 para o modelo 2, novamente, esse valor foi diluído nas interações que, no caso dessa variável, foram bastantes. No segundo modelo, note ainda, que o gênero, por si só, quase não tem impacto, sendo seu valor dominado pelas interações com a frequência de consumo de vegetais e prática de atividades físicas. No primeiro modelo, não inferimos nada do gênero, enquanto no segundo, há uma chance 90% menor de uma mulher estar em um estágio mais avançado de sobrepeso, se comparada a um homem com os mesmos outros valores nas variáveis. Nessa linha também, a frequência na prática de atividades físicas reduz, em 30% aproximadamente, nos dois modelos, a chance de estar em um estágio mais avançado de sobrepeso.

Ademais, falta falar sobre as interações colocadas no segundo modelo. Uma que não traz informações é a entre gênero e frequência de atividades físicas. A interação entre gênero e frequência consumo de verduras tende a aumentar as chances em 6 vezes, enquanto a interação entre o histórico familiar de sobrepeso e a frequência de consumo de vegetais aumentam em 3.7 essas chances de avançar para o próximo estado de sobrepeso.

Um fato interessante são as linhas de base, isto é, o intercepto. Sabemos que quanto maior for o intercept, mais efeitos não explicados pelas variáveis independentes estão sendo amortizados por ele e, como da para ver, no segundo modelo, essa quantidade diminui bastante, em comparação com o primeiro, onde a relação entre a primeira e a segunda classe, no caso, Insufficient Weight e Normal Weight mudou bastante, e conseguimos explicar melhor essa relação, enquanto as outras tendem a manter uma certa regularidade.

# 5 Conclusões

Como citado anteriormente, conseguimos responder as duas linhas gerais de perguntas propostas para este trabalho. No caso da segunda, o modelo se mostrou ineficiente em ser capaz de categorizar os indivíduos de maneira certa, ou seja, não há um poder preditivo com confiança alta. Já, para a primeira, com base no modelo estabelecido, conseguimos sim avaliar o impacto de determinadas condições, como a frequência de consumo hipercalórico de alimentos, no avanço do sobrepeso, medindo a incerteza de cada variável independente. Ainda falando dessa primeira pergunta, vemos que faz sentido algumas informações inferidas, como o histórico de sobrepeso na família impactar no peso da pessoa e a frequência de prática esportiva diminuir as chances de avançar para o próximo estágio de sobrepeso, por exemplo.

Em um primeiro momento, verificamos uma limitação no que diz respeito aos resultados obtidos. Isto é, talvez pelo conjunto de dados disponíveis ou alguma quebra de requisito, a capacidade preditiva, um dos objetivos, não foi conquistada. Nesse contexto, talvez utilizar o VIF (Variance Inflation Factor) para avaliar a multicolinearidade ou o uso de algum outro modelo poderia fazer não enfrentarmos essa limitação.

Com respeito as direções futuras, é fato que a adequação do modelo falhou no sentido de poder preditivo. Para tratar isso, uma linha futura de trabalhos poderia comparar este modelo com outras formas de categorização, por exemplo, árvores de decisão ou modelos onde a proporcionalidade de chances é rompida, assim a acurácia talvez seja maior e a resposta a segunda pergunta proposta fosse positiva. Concluindo então que, a categorização não é importante para condenar o sobrepeso, ao contrário dos estágios finais de obesidade, este pode aumentar a expectativa de vida [11], variando conforme grupo etário e, nesse sentido, seguir a linha de estudos desse trabalho, avaliando o impacto das variáveis independentes na categoria de sobrepeso, pode funcionar como um fator importante para o bem estar geral de uma sociedade onde o dito "terrorismo nutricional" [12] segue atuando.

# Referências


[1] V. S. dos Santos, "Obesidade." https://brasilescola.uol.com.br/saude/obesidade.htm. Acesso em 20 de junho de 2024.

[2] S. T. Nyberg, G. D. Batty, J. Pentti, M. Virtanen, L. Alfredsson, E. I. Fransson, M. Goldberg, K. Heikkilä, M. Jokela, A. Knutsson, M. Koskenvuo, T. Lallukka, C. Leineweber, J. V. Lindbohm, I. E. H. Madsen, L. L. Magnusson Hanson, M. Nordin, T. Oksanen, O. Pietiläinen, O. Rahkonen, R. Rugulies, M. J. Shipley, S. Stenholm, S. Suominen, T. Theorell, J. Vahtera, P. J. M. Westerholm, H. Westerlund, M. Zins, M. Hamer, A. Singh-Manoux, J. A. Bell, J. E. Ferrie, and M. Kivimäki, "Obesity and loss of disease-free years owing to major non-communicable diseases: a multicohort study," *The Lancet Public Health*, vol. 3, no. 10, pp. e490–e497, 2018.

[3] "Relatório vigitel 2023." https://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2023.pdf. Acesso em 20 de junho de 2024.

[4] "Estado nutricional vigitel 2006-2021." https://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2006-2021_estado_nutricional.pdf. Acesso em 20 de junho de 2024.

[5] A. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science, CRC Press, 2018.

[6] M. N. S. Abreu, "Uso de modelos de regressão logística ordinal em epidemiologia: um exemplo usando a qualidade de vida," 2007.

[7] J. P. Degani, "Análise de fatores que impactam no êxito de alunos em estado de vulnerabilidade social na unb." Trabalho de Conclusão de Curso (Bacharelado em Estatística), 2022. 64 f., il.

[8] F. M. Palechor and A. de la Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico," *Data in Brief*, vol. 25, p. 104344, 2019.

[9] L. S. Muhlheim, D. B. Allison, S. Heshka, and S. B. Heymsfield, "Do unsuccessful dieters intentionally underreport food intake?," *International Journal of Eating Disorders*, vol. 24, no. 3, pp. 259–266, 1998.

[10] T. Chang, N. Ravi, M. A. Plegue, K. R. Sonneville, and M. M. Davis, "Inadequate hydration, bmi, and obesity among us adults: NHANES 2009-2012," *Ann Fam Med*, vol. 14, pp. 320–324, July 2016. Erratum in: Ann Fam Med. 2020 Nov;18(6):485. doi: 10.1370/afm.2617.

[11] A. Chiolero, "Why causality, and not prediction, should guide obesity prevention policy," *The Lancet Public Health*, vol. 3, 09 2018.

[12] M. M. Rafael Figuereido, Christovão Paiva, "Terrorismo nutricional." https://www.arca.fiocruz.br/handle/icict/22827. Acesso em 23 de junho de 2024.