CIÊNCIA DE DADOS E ANÁLISE DE COMPONENTES PRINCIPAIS: UM ESTUDO SOBRE OBESIDADE NO BRASIL COM A BIBLIOTECA SCIKIT LEARN A PARTIR DA BASE DE DADOS VIGITEL

DATA SCIENCE AND PRINCIPAL COMPONENT ANALYSIS: A STUDY ON OBESITY IN BRAZIL USING THE SCIKIT-LEARN LIBRARY WITH THE VIGITEL DATABASE

ANDREY WATANUKI (FATEC RUBENS LARA)

andrey.watanuki@fatec.sp.gov.br

RENAN SANTOS (FATEC RUBENS LARA)

renan.santos138@fatec.sp.gov.br

RESUMO

Este estudo aborda a identificação de padrões do Índice de Massa Corporal (IMC) no Brasil entre 2006 e 2023, utilizando dados do sistema VIGITEL. Ao analisar os principais componentes. A análise sublinha a urgência de intervenções políticas e estratégias de saúde pública para combater essa escalada da obesidade e suas comorbidades. Este artigo oferece uma análise para formuladores de políticas, profissionais de saúde e pesquisadores interessados em abordar um dos desafios de saúde pública contemporâneos no Brasil.

Palavra-chaves: Análise de Principal Componente; Ciência de Dados; obesidade; Scikit-Learn; Vigitel.

ABSTRACT

This study addresses the identification of Body Mass Index (BMI) patterns in Brazil from 2006 to 2023, using data from the VIGITEL system. By analyzing the principal components, the analysis underscores the urgency of policy interventions and public health strategies to combat the rise in obesity and its comorbidities. This article provides an analysis for policymakers, health professionals, and researchers interested in addressing one of Brazil's contemporary public health challenges.

Keywords: *Principal Component Analysis; Data Science; obesity; Scikit-Learn; Vigitel.*

1 INTRODUÇÃO

A obesidade, caracterizada pelo acúmulo excessivo de gordura corporal, é uma condição crônica amplamente reconhecida pela Organização Mundial da Saúde (OMS, também conhecida por World Health Organization, WHO), diagnosticada por meio do Índice de Massa Corporal (IMC), calculado a partir da relação entre peso e altura. No Brasil, essa problemática ganha destaque como uma crescente preocupação de saúde pública, conforme indicado por dados da própria OMS. Precedente a 2019, o país testemunhou um aumento alarmante na prevalência da obesidade, com um crescimento significativo de 72%, elevando-se de 11,8% em 2006 para 20,3% em 2019. Esta tendência ascendente demanda medidas urgentes e eficazes para enfrentar essa questão de saúde pública (Abeso, 2024).

A pesquisa de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (VIGITEL) evidencia que a frequência da obesidade é similar entre homens e mulheres, mas tende a diminuir com o aumento do nível de escolaridade, sobretudo entre as mulheres (Abeso, 2024).

Adicionalmente, informações fornecidas pelo Ministério da Saúde e pela Organização Panamericana da Saúde ressaltam a preocupação com a obesidade infantil, afetando 12,9% das crianças brasileiras entre 5 e 9 anos e 7% dos adolescentes de 12 a 17 anos (OMS, 2024).

Ademais, de acordo com Holland *et al.* (2022), a conexão entre renda e padrões de consumo das famílias indica que maiores ganhos estão ligados a um consumo mais substancial, o que pode incluir alimentos mais calóricos. Contudo, níveis superiores de renda costumam estar relacionados à maior nível educacional, permitindo o acesso a informações e, por conseguinte, ações preventivas contra a obesidade. Este panorama reforça a necessidade de reconhecimento de padrões associados à obesidade e suas projeções, com o propósito de subsidiar a elaboração de políticas de saúde pública direcionadas e intervenções preventivas.

O objetivo deste estudo consiste em empregar análise do componente principal (PCA) para reduzir dimensionalidade e identificar padrões dos dados (Awari, 2024). Para este propósito, foram analisados os conjuntos de dados provenientes da VIGITEL, compreendendo o período de 2006 a 2023 (Vigitel, 2024). Tais conjuntos de dados oferecem informações atualizadas sobre a obesidade nas principais cidades do Brasil, viabilizando identificar as variáveis mais importantes.

No decorrer deste artigo, apresenta-se uma descrição dos procedimentos metodológicos utilizados no processo de aprendizado de máquina, abrangendo desde a fase inicial de aquisição e preparação de dados até a fase subsequente de desenvolvimento e aplicação. Especificamente, emprega-se a linguagem de programação Python e suas bibliotecas especializadas, como, ScikitLearn, Pandas e Matplotlib.

2 FUNDAMENTAÇÃO TEÓRICA

De acordo com o Ministério da Saúde (2024), a saúde vai além da simples ausência de enfermidades, incorporando-se o bem-estar integral, abrangendo aspectos físicos, mentais e sociais. Paralelamente, o conceito de qualidade de vida está intrinsecamente ligado, exigindo uma abordagem holística que leve em consideração não apenas o estado físico, mas também aspectos psicológicos e o contexto social do indivíduo. Conforme orientado pela OMS, a promoção da saúde engloba a adoção de hábitos alimentares saudáveis, a prática regular de exercícios físicos e a manutenção de uma hidratação adequada.

Participar ativamente de uma vida saudável não apenas previne enfermidades, mas também proporciona benefícios para o estado emocional, incluindo a redução de sintomas de ansiedade e depressão, além de favorecer um sono restaurador. A interação social e o autoconhecimento emergem como aspectos cruciais para a saúde mental, ressaltando-se a importância de relacionamentos saudáveis e práticas de autocuidado (Saúde, 2024).

As doenças crônicas não transmissíveis (DCNT) constituem um dos mais significativos desafios para a saúde pública tanto no Brasil quanto em nível global. Á vista disso, a proporção global de óbitos atribuíveis às DCNT aumentou para cerca de 74% em 2019 (Saúde, 2023; OMS, 2022).

Em 2019, o Brasil registrou 738.371 óbitos decorrentes de DCNT. Desses, 308.511 (41,8%) ocorreram de forma prematura, afetando indivíduos entre 30 e 69 anos. Isso resultou em uma taxa padronizada de mortalidade de 275,5 óbitos prematuros por 100 mil habitantes (Brasil, 2022).

2.1 Monitoramento da Saúde no Brasil

Sob a coordenação do Ministério da Saúde, a VIGITEL monitora a saúde da população brasileira através de entrevistas telefônicas, coletando informações sobre DCNT e seus fatores de risco.

Entre 2006 e 2019, ao longo das quatorze primeiras edições, a VIGITEL consolidou-se como o principal estudo de saúde do Brasil. Como parte de sua metodologia, foi estabelecido que o tamanho mínimo da amostra seria de aproximadamente 2 mil indivíduos por cidade. Esse critério foi adotado visando calcular a frequência de quaisquer indicadores na população adulta com um nível de confiança de 95% e uma margem de erro máxima de dois pontos percentuais. Além disso, para estimativas específicas por sexo, considera-se uma margem de erro máxima de três pontos percentuais, presumindo-se proporções semelhantes de homens e mulheres na amostra (OMS, 1991 *apud* Saúde, 2023).

Segundo a OMS (1991 *apud* Saúde, 2023), em localidades onde a cobertura de telefonia fixa era inferior a 40% dos domicílios e o número absoluto de domicílios com telefone não ultrapassa 50 mil, aceitavam-se amostras menores, variando entre 1.000 e 1.500 entrevistas. Nessas condições, as estimativas para a população adulta apresentaram um erro máximo de três pontos percentuais,

enquanto as estimativas específicas por gênero tiveram um erro máximo de quatro pontos percentuais.

De acordo com o Ministério da Saúde (2023), até a edição mais recente, foram conduzidas entrevistas com 305.682 homens e 500.487 mulheres, resultando na coleta de dados de um total de 806.169 residentes no Brasil.

No último relatório divulgado, foi constatado que a prevalência da obesidade entre a população adulta no Brasil aumentou significativamente. Enquanto em 2021 o fenômeno afetava em média 19,8% dos adultos, em 2023 esse número saltou para um quarto da população acima dos 18 anos. Este aumento alarmante reflete uma tendência de crescimento da obesidade no país, evidenciando a necessidade de medidas para lidar com esse problema de saúde pública.

2.2 Obesidade e Saúde no Brasil

A condição de obesidade é identificada pelo sobrepeso resultante do acúmulo de gordura corporal, que é indicado por um IMC igual ou superior a 30, que foi proposto por Adolphe Quetelet em seu livro "*Sur l'Homme Et Le Développement de Ses Facultés, Ou Essai de Physique Sociale*", publicado em 1835. Durante muitos anos, esse índice foi conhecido como Índice de Quetelet, em homenagem ao seu criador. (Quetelet, 1835 *apud* Magosso, 2023). kg/m^2 é a unidade de medida de IMC (Saúde, 2024). É importante ressaltar que a obesidade é considerada uma doença crônica, que por si só pode desencadear uma série de outras condições adversas, desde problemas cardiovasculares a diabetes, conhecidos como comorbidades pelos cientistas (Abeso, 2024).

No Brasil, em 2015, foram registradas 116.976 mortes devido às doenças causadas pela obesidade. Isso significa 2,44 vezes mais mortes por excesso de peso do que por assassinatos. Em Barbados, Uruguai, Chile, Cuba e Argentina, os casos superam entre 10 e 19 vezes os causados pela violência (Berdegué; Aguirre, 2018).

A proporção de obesos na população com 20 anos ou mais de idade mais que dobrou no país entre 2003 e 2019, passando de 12,2% para 26,8%. Nesse período, a obesidade feminina subiu de 14,5% para 30,2% , enquanto a obesidade masculina passou de 9,6% para 22,8% (Brasil, 2020).

Entre 2019 e 2023, foram registradas 40.900 internações com obesidade como diagnóstico principal. Embora a prevalência de pacientes hospitalizados por obesidade no Brasil tenha reduzido de 36% em 2019 para 23% em 2023, essa diminuição explica o fato de que os anos de 2020 e 2021 registraram o menor número de internações. Esse período coincidiu com a pandemia de COVID-19 no Brasil, resultando em uma redução das hospitalizações por obesidade. (Costa *et al.*,2024).

Em comparação aos resultados citados por Berdegué e Aguirre (2018), bem como o crescimento observado entre os anos de 2003 e 2019, o panorama no Brasil é preocupante. A análise preditiva baseada em séries temporais e a progressão da obesidade pode revelar uma tendência alarmante na área da saúde.

Esses dados forneceram bases confiáveis a fim de nortear as políticas públicas na área de saúde contribuindo para identificar problemas e controlar o crescimento das condições crônicas que geram as comorbidades em geral.

3 PROCEDIMENTOS METODOLÓGICOS

Este artigo é um estudo de corte transversal e consiste em analisar o aumento da prevalência de obesidade, mediante a análise de conjuntos de dados segmentados por variáveis como faixa etária, localização geográfica, etnia, nível educacional, prática de atividade física e estado de saúde, em relação ao IMC. Os dados são examinados para constatar o aumento da obesidade, empregando estudos da Ciência de Dados, empregando técnicas de estatística descritiva com base nos registros da VIGITEL.

Todos os dados empregados neste estudo foram obtidos no repositório de arquivos da VIGITEL. Foram baixados um total de 18 arquivos, incluindo 16 arquivos de dados no formato XLS, 1 arquivo de dados no formato XLSX e 1 arquivo de dicionário no formato XLS.

As variáveis mais relevantes do conjunto de dados são analisadas, juntamente com a aplicação da Análise de Componentes Principais (PCA) para a redução da dimensionalidade dos dados, considerando as diferentes faixas de IMC.

A técnica de PCA é um método estatístico utilizado para reduzir a dimensionalidade de conjuntos de dados complexos. Utilizando Python, essa técnica pode ser implementada de maneira eficiente, proporcionando insights valiosos a partir dos dados. A PCA transforma os dados originais em componentes principais, que são combinações lineares das variáveis originais, selecionadas para maximizar a variância dos dados, permitindo capturar a maior quantidade de informação com o menor número de componentes.

Para implementar a PCA com Python, é necessário seguir alguns passos: importação das bibliotecas necessárias, preparação e padronização dos dados, aplicação da PCA, análise dos resultados, interpretação dos componentes principais e utilização desses componentes. A PCA é amplamente utilizada em estatística, ciência de dados e aprendizado de máquina, devido à sua capacidade de simplificar a análise de dados complexos enquanto mantém a maior quantidade de informação possível.

3.1 Ferramenta da Área de Ciência de Dados

As ferramentas apresentadas nesta seção abrangem linguagens de programação e bibliotecas especializadas, essenciais para a análise, manipulação e visualização de dados.

a) Python

Optou-se pelo uso da linguagem de programação Python devido à sua capacidade de empregar bibliotecas que facilitam a compreensão das análises. As características fundamentais desta linguagem incluem sua natureza interpretada,

interativa e orientada a objetos. Ela incorpora conceitos como módulos, exceções, tipagem dinâmica, tipos de dados dinâmicos de alto nível e a definição de classes. Ademais, Python suporta diversos paradigmas de programação, incluindo o procedural e o funcional. A linguagem combina potência com uma sintaxe extremamente clara, além de oferecer interfaces para uma variedade de chamadas de sistema e bibliotecas (Python, 2024).

Segundo Python (2024), disponibiliza-se o módulo *os*. Este módulo oferece uma maneira portátil de utilizar funcionalidades dependentes do sistema operacional, da mesma forma, possibilita a abertura de arquivos e sua conversão em variáveis.

b) Pandas

O Pandas é uma biblioteca Python que oferece estruturas de dados rápidas, flexíveis e expressivas, projetadas para facilitar o trabalho com dados "relacionais" ou "rotulados". Seu objetivo é ser o componente fundamental de alto nível para a análise prática e real de dados em Python. Além disso, busca se tornar a ferramenta de código aberto mais poderosa e flexível para análise e manipulação de dados disponível em qualquer linguagem (Pandas, 2024).

Além disso, o Pandas oferece funcionalidades avançadas para análise e visualização de dados. Isso inclui a capacidade de explorar os dados em busca de padrões e além de criar tabelas, gráficos e diagramas para visualizar os dados de forma clara e intuitiva (Gaspar *et al.,* 2023).

A biblioteca pandas foi utilizada para carregar e manipular os dados, uma vez que dispõe de funções que permitem a leitura de arquivos Excel Spreadsheet (xls) e Excel Open Xml Spreadsheet (xlsx), convertendo-os em *dataframes* para posterior tratamento, limpeza e revisão (Pandas, 2024).

d) Scikit-learn

O Scikit-learn é um módulo Python que uma ampla variedade de algoritmos de aprendizado de máquina de última geração incorpora para resolver problemas de média escala, tanto supervisionados quanto não supervisionados (PEDROGOSA et al., 2011).

Hao e Ho (2019) esclarece que o aprendizado supervisionado é um conjunto específico de algoritmos de machine learning que estabelecem uma relação entre as variáveis de características e suas variáveis-alvo correspondentes. Para utilizar métodos de aprendizado supervisionado, é necessário que tanto as características quanto seus respectivos rótulos sejam conhecidos. Esse tipo de aprendizado pode ser dividido em duas categorias dependendo da natureza dos rótulos: regressão, quando os rótulos são contínuos, e classificação, quando os rótulos são discretos.

Já os modelos de aprendizado de máquina são frequentemente suscetíveis ao overfitting. Por isso, uma estratégia comum para avaliar o desempenho de um modelo em problemas de classificação é a técnica de validação cruzada. Nesse método, os dados são divididos em múltiplas partes de forma aleatória, sendo

algumas utilizadas como conjunto de treinamento e outras como conjunto de validação (HAO; HO, 2019).

Entende-se que ao demonstrar a transformação de dados, o aprendizado supervisionado e a avaliação de modelos, no qual, devem atender aos nossos objetivos de fornecer informações preditivas esperadas, demonstra-se o necessário para o entendimento sobre as técnicas para o desenvolvimento desse estudo.

e) Matplotlib

Matplotlib é um pacote Python utilizado para plotagem, capaz de gerar gráficos de alta qualidade para produção. Foi desenvolvido para criar tanto gráficos simples quanto complexos com apenas alguns comandos. Matplotlib é construído para utilizar NumPy e outras extensões de código. Com algumas linhas de código, é possível criar uma variedade de gráficos, incluindo histogramas, espectros de potência, gráficos de barras, gráficos de erros, gráficos de dispersão, entre outros. Ele opera através de uma interface processual chamada "pylab", que é baseada em uma máquina de estados. Além disso, Matplotlib proporciona gráficos com qualidade de publicação e saída em postscript, tornando-os adequados para inclusão em documentos TeX. Ele também pode ser integrado em interfaces gráficas do usuário para o desenvolvimento de aplicativos (ARI; USTAZHANOV, 2014).

De acordo com Matplotlib (2023, apud GASPAR et al., 2023) a Matplotlib possibilita a customização dos gráficos, permitindo a modificação de cores, inserção de legendas, títulos e outros elementos. Essa versatilidade viabiliza criação de visualizações adaptadas para diferentes tipos de dados e análises, assegurando uma representação visual precisa e informativa dos resultados obtidos.

Além disso, Matplotlib (2023, apud GASPAR et al., 2023) elucida que a biblioteca é capaz de suportar múltiplos subplots em um único gráfico, o que permite a visualização de diferentes conjuntos de dados lado a lado, facilitando a comparação e análise.

3.2 Tratamento de Dados – ETL

O processo de Extract Transform Load (ETL) destaca-se como uma das abordagens mais eficazes e amplamente adotadas para a integração de dados. Este método compreende três etapas principais: extração, transformação e carregamento de dados (Nwokeji; Matovu, 2021).

a) Extração

Foram selecionados os arquivos que continham todos os dados disponíveis no repositório online da VIGETEL, separados por ano, juntamente com seus respectivos dicionários. Foi excluída da seleção os arquivos "Dicionario-de-dados-Vigitel-PopNegra.xls" e "Vigitel-2018-PopNegra.xls", devido à falta de consistência anual em sua disponibilidade.

Todas as planilhas baixadas foram disponibilizadas no formato Excel Spreadsheet (XLS) e Excel Open Xml Spreadsheet (XLSX).

Ao compreender a organização dos dados, identificam-se informações relevantes para o escopo do estudo, dando início à fase de transformação dos dados.

b) Transformação

Durante a fase inicial de carregamento dos dados utilizando o módulo os, observa-se uma variação nas extensões dos arquivos de planilha. Diante dessa observação, foi estabelecida uma condição para garantir a consistência no processo de concatenação de todas as planilhas usando a biblioteca Pandas mediante ao Python. Essa condição foi especificamente definida para identificar e carregar o formato dos arquivos encontrados na pasta, seja em formato XLS ou XLSX, a fim de assegurar um carregamento adequado. Os dados foram então armazenados em uma variável denominada "pesquisa", representada como um *dataframe*.

Após a conclusão do carregamento dos dados, as colunas relevantes para o escopo deste artigo foram renomeadas de acordo com as especificidades da análise. A Tabela 1 apresenta os nomes originais das colunas e os seus respectivos nomes após a transformação.

Tabela 1 – Nomes das Colunas – Antes e Depois da Renomeação

Nomes Originais	**Nomes Transformados**
ano	**Ano**
cidade	**Cidade**
fet	**Faixa Etária**
fesc	**Faixa de Escolaridade**
imc	**IMC**
q7	**Genero**
q42	**Exercicio Fisico**
q74	**Estado de Saúde**
q69	**Cor**

Fonte: Adaptado de Vigitel (2024)

Após a conclusão da renomeação de todas as colunas conforme exibido na Tabela 1, o próximo passo foi realizar a filtragem dos dados. Este processo envolveu a substituição do *dataframe* anteriormente criado, denominado "pesquisa", por uma seleção contendo apenas os dados renomeados.

Para facilitar a compreensão dos procedimentos subsequentes, é importante destacar que a maioria das perguntas feitas pela VIGITEL são de múltipla escolha, seguindo um padrão de respostas específicas, tais como "Gênero", "Faixa de Escolaridade", "Exercício Físico", "Estado de Saúde", "Cor" e "Cidade". As respostas dos participantes foram armazenadas como números e associadas a

seu dicionário. Ao carregar os dados pelo Pandas, observou-se a necessidade de converter as respostas numéricas de *float* para *int*. Dessa forma foram realizados procedimentos utilizando funções criadas pelos autores, que associaram os números presentes nas planilhas às respostas correspondentes, substituindo os valores.

Nesta etapa, também foi necessário verificar a existência de duplicidade e de inconsistências nos dados, além de filtrar as respostas selecionadas, tais como: "não sabia responder", "não queria informar", "não lembrava" e "não quis responder", conforme definido no dicionário.

Após a substituição das respostas, foi criada uma nova coluna denominada "Faixa IMC", que relaciona o IMC com faixas específicas de IMC, tais como "Abaixo do peso", "Peso normal", "Sobrepeso", "Obesidade grau 1", "Obesidade grau 2" e "Obesidade grau 3", com as respectivas faixas de IMC associadas (Abeso, 2024).

Na etapa seguinte, foi essencial proceder com a conversão de cada componente a ser analisado em vetores.

c) Carregamento

Foi essencial empregar a biblioteca Matplotlib para a visualização dos dados em cada conjunto de variáveis submetido à análise.

4 ANÁLISE E RESULTADO

A PCA foi realizada com o objetivo de identificar padrões e reduzir a dimensionalidade dos dados. A matriz de covariância dos dados foi calculada e a decomposição revelou os autovalores e autovetores associados, que representam a variância explicada e a direção dos componentes principais, respectivamente.

a) Matrix de Covariância

A matriz de covariância fornecida quantifica a variação conjunta entre pares de variáveis em um conjunto de dados. Cada entrada na matriz de covariância indica a magnitude e a direção do relacionamento linear entre duas variáveis específicas. Aqui está a Matriz Quadrada Reduzida da Matriz de Covariância, conforme apresentado abaixo:

$$\begin{bmatrix} 60.9469866 & 0.0506822 & 0.0101405 & 0.1173670 & 0.0305347 & 0.9203965 & -0.0247875 \\ 0.0506822 & 2.6890467 & 0.0623525 & -0.3139765 & 0.0559811 & 4.6391445 & -0.2041929 \\ 0.0101405 & 0.0623525 & 0.2402770 & 0.0006623 & 0.0287440 & 0.0598135 & -0.0169334 \\ 0.1173670 & -0.3139765 & 0.0006623 & 0.6023197 & -0.0748594 & -1.8638820 & -0.0853671 \\ 0.0305347 & 0.0559811 & 0.0287440 & -0.0748594 & 0.2486647 & 0.6699693 & 0.0155765 \\ 0.9203965 & 4.6391445 & 0.0598135 & -1.8638820 & 0.6699693 & 2893.5066 & -0.4205705 \\ -0.0247875 & -0.2041929 & -0.0169334 & -0.0853671 & 0.0155765 & -0.4205705 & 1.7106209 \end{bmatrix}$$

b) Autovalores e Autovetores

Os autovalores de uma matriz de covariância são números que descrevem a variabilidade dos dados ao longo das direções principais dos dados. Na PCA, a matriz de covariância é construída a partir dos dados originais e desempenha um papel crucial na identificação dessas direções principais, também conhecidas como componentes principais. Aqui estão os detalhes sobre os dois maiores:

Autovalores:

$$\begin{bmatrix} 2893.5157 \\ 60.9469 \\ 2.7643 \\ 1.6885 \\ 0.5569 \\ 0.2073 \\ 0.2645 \end{bmatrix}$$

Autovetores:

$$\begin{bmatrix} -3.2893 \times 10^{-4} & -9.9999 \times 10^{-1} & -6.6363 \times 10^{-4} & 5.2039 \times 10^{-4} & -1.8784 \times 10^{-3} & 6.9495 \times 10^{-4} & 5.6034 \times 10^{-4} \\ -1.6054 \times 10^{-3} & -8.5410 \times 10^{-4} & 9.7429 \times 10^{-1} & 1.6088 \times 10^{-1} & 1.5553 \times 10^{-1} & 1.6562 \times 10^{-2} & -1.9840 \times 10^{-2} \\ -2.0798 \times 10^{-5} & -1.7262 \times 10^{-4} & 2.5491 \times 10^{-2} & -4.1924 \times 10^{-3} & 9.2189 \times 10^{-3} & -7.3887 \times 10^{-1} & 6.7328 \times 10^{-1} \\ 6.4510 \times 10^{-4} & -1.9609 \times 10^{-3} & -1.3386 \times 10^{-1} & -1.2487 \times 10^{-1} & 9.6214 \times 10^{-1} & -1.3150 \times 10^{-1} & -1.5320 \times 10^{-1} \\ -2.3233 \times 10^{-4} & -4.9447 \times 10^{-4} & 2.4432 \times 10^{-2} & 2.3273 \times 10^{-2} & -1.9900 \times 10^{-1} & -6.6064 \times 10^{-1} & -7.2305 \times 10^{-1} \\ -9.9999 \times 10^{-1} & 3.2921 \times 10^{-4} & -1.6820 \times 10^{-3} & -2.0370 \times 10^{-4} & 4.3229 \times 10^{-4} & 5.6204 \times 10^{-5} & 8.7368 \times 10^{-5} \\ 1.4368 \times 10^{-4} & 4.3297 \times 10^{-4} & -1.7770 \times 10^{-1} & 9.7875 \times 10^{-1} & 1.0196 \times 10^{-1} & -6.9573 \times 10^{-3} & 3.7915 \times 10^{-3} \end{bmatrix}$$

Dois maiores autovalores:

$$\begin{bmatrix} 2893.5157 \\ 60.9469 \end{bmatrix}$$

Dois maiores autovetores:

$$\begin{bmatrix} -3.2493 \times 10^{-4} & -1.6048 \times 10^{-3} & -2.0711 \times 10^{-5} & 6.4445 \times 10^{-4} & -2.3161 \times 10^{-4} & -9.9999 \times 10^{-1} & 1.4553 \times 10^{-4} \\ -9.9999 \times 10^{-1} & -8.3568 \times 10^{-4} & -1.6795 \times 10^{-4} & -1.9506 \times 10^{-3} & -4.9780 \times 10^{-4} & 3.2520 \times 10^{-4} & 4.2175 \times 10^{-4} \end{bmatrix}$$

A interpretação da análise revela que o primeiro autovalor (2893.5157676781123 2893.5157676781123) é substancialmente maior do que os outros, indicando que o primeiro componente principal explica a maior parte da variância dos dados. O segundo autovalor (60.94698589168302 60.94698589168302) também é significativo, sugerindo que o segundo componente principal captura uma parte substancial da variância remanescente.

Os autovetores fornecem informações sobre quais variáveis contribuem mais para cada componente principal. O autovetor correspondente ao maior autovalor (-3.24938841e-04, -1.60486959e-03, -2.07119808e-05, 6.44454023e-04, -2.31611708e-04, -9.99998414e-01, 1.45531020e-04) mostra que a sexta variável tem a maior contribuição para a primeira componente principal, dada a magnitude

do coeficiente (-9.99998414e-01). O autovetor correspondente ao segundo maior autovalor (-9.99997469e-01, -8.35687981e-04, -1.67952813e-04, -1.95061734e-03, -4.97801619e-04, 3.25202774e-04, 4.21750016e-04) indica que a primeira variável tem a maior contribuição para a segunda componente principal, dada a magnitude do coeficiente (-9.99997469e-01).

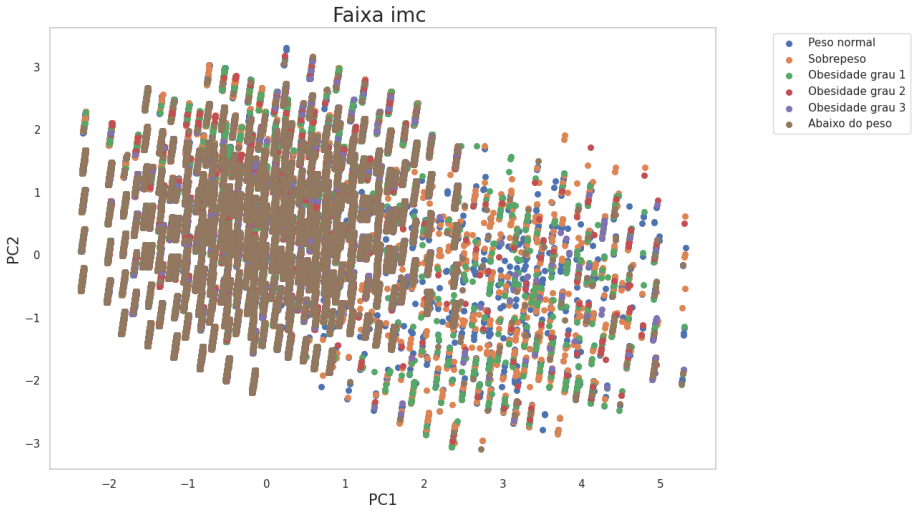
A análise PCA indica que a maior parte da variação nos dados é explicada pelo primeiro componente principal, que é fortemente influenciada pela sexta variável original. O segundo componente principal também contribui significativamente para a variação total e é principalmente influenciada pela primeira variável original. Portanto, ao reduzir a dimensionalidade dos dados, as variáveis originais seis e um são as mais importantes e capturam a maior parte da variação no conjunto de dados.

Primeiro Componente Principal: Este componente captura a maior parte da variabilidade dos dados, predominantemente influenciado pela variável "Abaixo do Peso". A magnitude do coeficiente do "Estado de Saúde" no autovetor correspondente ao maior autovalor indica que esta variável é a principal responsável pela variação explicada por este componente. Em termos práticos, isso significa que diferenças nas Faixa de IMC são os principais fatores de variação nos dados analisados. Portanto, ao reduzir a dimensionalidade dos dados, o "Abaixo do Peso" surge como a variável mais importante.

Segundo Componente Principal: O segundo componente principal, embora também influenciado pela variável "Abaixo do Peso", captura variações menores associadas a outras variáveis, como "Peso Normal" e "Sobrepeso". Este componente ainda é significativo, mas explica menos variância que o primeiro. A presença dessas variáveis indica que, além do estado de saúde, há outros fatores que contribuem para a variabilidade nos dados.

c) Resultado

A plotagem dos dados no espaço dos componentes principais revela a distribuição e a concentração dos pontos correspondentes às diferentes variáveis. Observou-se que os pontos relacionados a determinadas faixas de "Abaixo do Peso" tendem a se agrupar, indicando uma clara separação baseada nesta variável. As variações capturadas pelo segundo componente principal mostram uma dispersão maior, sugerindo que "Peso Normal" e "Sobrepeso" adicionam nuances à análise.



5 CONSIDERAÇÕES FINAIS

A análise PCA realizada neste estudo revelou que a maior parte da variação nos dados é explicada pela variável "Abaixo do Peso". Isso sugere que, ao analisar e interpretar os dados, esta variável deve ser considerada como a mais influente. A segunda componente principal também é relevante, destacando a importância de outras variáveis como "Peso Normal" e "Sobrepeso". A metodologia de PCA foi eficaz para simplificar a interpretação dos dados, destacando as principais direções de variação e ajudando na identificação de características fundamentais que diferenciam os estados de saúde nas diversas cidades e faixas de gênero.

Essa abordagem facilita a tomada de decisões informadas no contexto de análise de saúde pública, permitindo uma melhor compreensão das influências sobre o estado de saúde da população. A visualização simplificada proporcionada pela PCA oferece uma representação clara dos dados, auxiliando na identificação de padrões e tendências que poderiam não ser evidentes em uma análise multivariada mais complexa.

REFERÊNCIAS

Código-Fonte no Github
https://github.com/renanmartinssantos/ObsVigiPCA

Base de Dados no Site da Vigente
https://svs.aids.gov.br/download/Vigitel

ABESO – Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica. **Mapa da Obesidade**. Disponível em: https://abeso.org.br/obesidade-e-sindrome-metabolica/mapa-da-obesidade/. Acesso em: 06 mar. 2024.

ABESO – Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica. **Obesidade e sobrepeso**. Disponível em: https://abeso.org.br/conceitos/obesidade-e-sobrepeso/. Acesso em: 06 abr. 2024.

ABESO – Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica. **Calculadora de IMC**. Disponível em: https://abeso.org.br/obesidade-e-sindrome-metabolica/calculadora-imc/. Acesso em: 20 maio 2024.

AMORIM, Guilherme; LERNER, Julianna; FREITAS, Natália. **Análise e Previsão de Comportamento de Ações da B3 utilizando Python e Séries Temporais**. 3º Encontro de Ciência de Dados. 2023. Disponível em: https://www.even3.com.br/iii-encontro-de-ciencia-de-dados-das-fatecs-380931(https://www.youtube.com/watch?v=u57sdMkSNMo). Acesso em: 05 maio 2024.

BERDEGUÉ, Julio; AGUIRRE, Pablo. **Obesidade que mata**. 2018. Disponível em: https://oglobo.globo.com/opiniao/obesidade-que-mata-22386691. Acesso em: 08 maio 2024.

BRASIL. Ministério da Saúde. **Saúde apresenta atual cenário das doenças não transmissíveis no Brasil**. Brasília, DF: MS, 2022. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/doencas-cronicas-nao-transmissiveis-dcnt/09-plano-de-dant-2022_2030.pdf. Acesso em: 19 maio 2024.

BRASIL, Governo do. **Pesquisa do IBGE mostra aumento da obesidade entre adultos**. 2020. SAÚDE. Disponível em: https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/10/pesquisa-do-ibge-mostra-aumento-da-obesidade-entre-adultos. Acesso em: 19 maio 2024.

CASTRO, Guilherme. **Modelagem Bayesiana De Dados De Degradação: Um Estudo Comparativo Dos Softwares Jags E Stan**. 2022. 52 f. Monografia (Especialização) — Curso de Estatística, Universidade Federal De Minas Gerais, Belo Horizonte, 2022. Disponível em: https://repositorio.ufmg.br/bitstream/1843/53870/1/Monografia_Guilherme_Hoffman.pdf. Acesso em: 29 maio 2024.

COSTA, Maira Damasceno; COSTA, Maiane Damasceno; NEVES, Lucas Mendes Fagundes; RODRIGUES, Ester Martins França; SILVA, Janaína Inácio da;

MAGALHÃES, Elizandro Mesquita; MARTINS, Maiara de Souza; SILVA, Bruna Figueredo Valadão da; MIRANDA, Yara Farias; NASCIMENTO, Jaine de Andrade do. PANORAMA DA MORBIDADE HOSPITALAR DE PACIENTES INTERNADOS COM OBESIDADE NO BRASIL. **Brazilian Journal Of Implantology And Health Sciences**, [S.L.], v. 6, n. 5, p. 838-848, 11 maio 2024. Brazilian Journal of Implantology and Health Sciences. http://dx.doi.org/10.36557/2674-8169.2024v6n5p838-848. Acesso em: 19 maio 2024

FÁVARO, Thatiana Regina; SANTOS, Ricardo Ventura; CUNHA, Geraldo Marcelo da; LEITE, Iuri da Costa; COIMBRA JUNIOR, Carlos E. A.. Obesidade e excesso de peso em adultos indígenas Xukuru do Ororubá, Pernambuco, Brasil: magnitude, fatores socioeconômicos e demográficos associados. **Cadernos de Saúde Pública**, [S.L.], v. 31, n. 8, p. 1685-1697, ago. 2015. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/0102-311x00086014.

GARLAPATI, A.; KRISHNA, D. R.; GARLAPATI, K.; SRIKARA Y. N. mani; RAHUL, U.; NARAYANAN, G. **Stock Price Prediction Using Facebook Prophet and Arima Models**. 2021. Disponível em: https://ieeexplore.ieee.org/document/9418057. Acesso em: 08 maio 2024.

GASPAR, Juliano de Souza; REIS, Zilma Silveira Nogueira; OLIVEIRA, Isaias José Ramos de; SILVA, Ana Paula Couto da; DIAS, Cristiane dos Santos. **Introdução à Análise de Dados em Saúde com Python**. 25 abr. 2023. Zenodo. Disponível em: https://zenodo.org/records/7865448. Acesso em: 05 abr. 2024.

MAGOSSO, R. (2023). **Índice de Massa Corporal: de volta à origem**. Intercursos Revista Científica, 22(2), 2–4. Disponível em: https://revista.uemg.br/index.php/intercursosrevistacientifica/article/view/8303. Acesso em: 20 maio 2024.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Técnicas de pesquisa: planejamento e execução de pesquisa, amostragens e técnicas de pesquisa, elaboração, análise e interpretação de dados**. 9. ed. Rio de Janeiro: Atlas, 2021.

MEDRI, Dr. Waldir. **Análise Exploratória de Dados**. 2011. 82 f. Tese (Doutorado) - Curso de Curso de Especialização "Lato Sensu" em Estatística, Universidade Estadual de Londrina, Londrina, 2011. Acesso em: 06 abr. 2024.

MELANI, Arthur Henrique de Andrade. System fault diagnosis based on Bayesian networks and SysML. 2020. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 2020. Disponível em: https://www.teses.usp.br/teses/disponiveis/3/3151/tde-31032021-162257/. Acesso em: 29 maio 2024.

NWOKEJI, Joshua C.; MATOVU, Richard. A Systematic Literature Review on Big Data Extraction, Transformation and Loading (ETL). **Lecture Notes In Networks And Systems**, [S.L.], p. 308-324, 2021. Springer International

Publishing. http://dx.doi.org/10.1007/978-3-030-80126-7_24. Acesso em: 29 maio 2024

PANDAS. **How do I select a subset of a DataFrame?**. 2024. Disponível em: https://pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html. Acesso em: 20 maio 2024.

SAÚDE, Ministério da. **Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis**. 2024. VIGITEL. Disponível em: https://svs.aids.gov.br/download/Vigitel. Acesso em: 05 maio 2024.

SAÚDE, Ministério da. **O que significa ter saúde?**. Muito além da ausência de doenças, é preciso considerar o bem-estar físico, mental e social. 2021. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-brasil/eu-quero-me-exercitar/noticias/2021/o-que-significa-ter-saude. Acesso em: 08 maio 2024.

SAÚDE, Ministério da. **Vigilância De Fatores De Risco E Proteção Para Doenças Crônicas Por Inquérito Telefônico**. 2006. Estimativas Sobre Freqüência E Distribuição Sócio-Demográfica De Fatores De Risco E Proteção Para Doenças Crônicas Nas Capitais Dos 26 Estados Brasileiros E No Distrito Federal Em 2006. Disponível em: hhttps://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2006.pdf. Acesso em: 08 jun. 2024.

SAÚDE, Ministério da. **04/3 – Dia Mundial da Obesidade. 2022**. Disponível em: https://bvsms.saude.gov.br/04-3-dia-mundial-da-obesidade/. Acesso em: 08 jun. 2024.

SAÚDE, Ministério da. **Vigilância De Fatores De Risco E Proteção Para Doenças Crônicas Por Inquérito Telefônico**. 2007. Estimativas Sobre Freqüência E Distribuição Sócio-Demográfica De Fatores De Risco E Proteção Para Doenças Crônicas Nas Capitais Dos 26 Estados Brasileiros E No Distrito Federal Em 2007. Disponível em: hhttps://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2007.pdf. Acesso em: 08 jun. 2024.

SAÚDE, Ministério da. **VIGITEL: sistema de vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico**. 2023. Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/vigitel/vigitel-brasil-2023-vigilancia-de-fatores-de-risco-e-protecao-para-doencas-cronicas-por-inquerito-telefonico. Acesso em: 06 abr. 2024.

SEABORN. **Seaborn: statistical data visualization**. 2024. Disponível em: https://seaborn.pydata.org/index.html. Acesso em: 20 maio 2024.

SILVA, André Luiz Carvalhal da. **Introdução à análise de dados**. Rio de Janeiro: E-Papers, 2009. 158 p. Acesso em: 06 abr. 2024.

SOARES, Josiel Ferreira. **Análise preditiva: Importância e vantagens para os negócios**. 2018. 11 f. TCC (Graduação) - Curso de Especialização em Governança da Tecnologia da Informação, Universidade do Sul de Santa Catarina, Santa Catarina, 2017. Acesso em: 08 maio 2024.

TAYLOR, Sean J; LETHAM, Benjamin. **Forecasting at Scale**. 2017. Disponível em: https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale. Acesso em: 09 maio 2024.

OMS – Organização Mundial da Saúde. **Obesity**. 2024. Disponível em: https://www.who.int/health-topics/obesity#tab=tab_1. Acesso em: 06 mar. 2024.

OMS – Organização Mundial da Saúde. **World Health Statistics**. Disponível em: https://iris.who.int/bitstream/handle/10665/272596/9789241565585-eng.pdf?ua=1. Acesso em: 19 maio 2024.

ZORZETTO, Ricardo. **Linhagens da variante ômicron elevam o total de casos de Covid-19 no final de 2022**. 2023. Disponível em: https://revistapesquisa.fapesp.br/linhagens-da-variante-omicron-elevam-o-total-de-casos-de-covid-19-no-final-de-2022/. Acesso em: 08 jun. 2024