Relatório Interpretativo do Projeto de Análise Exploratória de Dados

Introdução

Este relatório apresenta uma Análise Exploratória de Dados (EDA) conduzida sobre o conjunto de dados "Dados de Saúde e Hábitos de Vida". O objetivo do projeto é utilizar a linguagem R para explorar informações sobre saúde e estilo de vida de indivíduos, identificar padrões, gerar insights e formular hipóteses. A análise foi orientada pelas seguintes perguntas de pesquisa, que foram testadas estatisticamente ao longo do processo:

- Há diferenças significativas no Índice de Massa Corporal (IMC) entre homens e mulheres?
- Existe relação entre o nível de atividade física e a pressão sistólica?
- A idade influencia no histórico de doenças?
- Fumantes apresentam maior prevalência de doenças crônicas?
- Há uma relação entre a quantidade de horas de sono e a pressão sistólica?

Metodologia

A análise foi conduzida seguindo as etapas de tratamento de dados e análise exploratória, conforme as especificações do projeto.

Limpeza e Tratamento de Dados:

- Identificação de NAs: Foram quantificados os valores ausentes nas variáveis, como Pressao_Sistolica, Pressao_Diastolica, Fumante, Historico_Doenca e Consumo_Alcool_semanal_ml.
- Padronização de Categóricas: A variável Sexo foi padronizada para "Feminino" e "Masculino" e a coluna Historico_Doenca foi padronizada para "Sim" e "Não".
- Tratamento de Dados Inconsistentes: Valores com formato incorreto em Horas_Sono e Consumo_Alcool_semanal_ml foram convertidos para NA antes do tratamento.
- Conversão de Datas: A coluna Data_Exame foi convertida para o formato de data.
- Imputação de NAs: Para variáveis numéricas, foi utilizada a mediana; para variáveis categóricas, a moda.

Engenharia de Variáveis:

- Foi criada uma nova variável, o Índice de Massa Corporal (IMC), calculado por: IMC = Peso (kg) / (Altura (m))²

Análise Exploratória:

- Análises Univariadas: Foram gerados gráficos de distribuição (histogramas e gráficos de pizza) para variáveis como Idade, IMC, Pressao_Sistolica, Sexo e Fumante.

- Análises Bivariadas: Foram exploradas relações entre pares de variáveis, incluindo Numérica vs. Numérica, Numérica vs. Categórica e Categórica vs. Categórica.

Resultados e Análises

IMC entre Homens e Mulheres Hipótese nula (H0): Existe diferença significativa na média de IMC entre homens e mulheres. Achado: p-valor = 0.9008 Conclusão: Não há evidência para rejeitar a hipótese nula. O gênero não parece influenciar o IMC nessa amostra.

Atividade Física e IMC Hipótese nula (H0): Existe diferença nas médias de IMC entre os níveis de atividade física. Achado: p-valor = 0.48683 Conclusão: Não foi encontrada diferença significativa. O nível de atividade física não influenciou o IMC na amostra.

Idade e Histórico de Doenças Hipótese nula (H0): A idade média é igual entre pessoas com e sem histórico de doenças. Achado: p-valor = 0.3692 Conclusão: Não há evidência para rejeitar a hipótese nula. A idade média entre os grupos não difere significativamente.

Tabagismo e Histórico de Doenças Hipótese nula (H0): Ser fumante e ter doenças crônicas são eventos independentes. Achado: p-valor = 0.8901 Conclusão: Não há associação estatisticamente significativa entre fumar e histórico de doenças crônicas.

Horas de Sono e Pressão Arterial Hipótese: Existe correlação entre horas de sono e pressão sistólica. Achado: A análise não revelou correlação linear significativa. Conclusão: Relação não linear. Outros fatores (como estresse ou qualidade do sono) podem ser mais relevantes.

Conclusão

A análise revelou que os hábitos avaliados — tabagismo, atividade física, idade e horas de sono — não apresentaram relações estatisticamente significativas com os indicadores de saúde analisados (IMC e pressão sistólica).

Principais conclusões:

- Gênero não influencia o IMC de forma significativa.
- Atividade física e idade não demonstraram relação direta com IMC ou histórico de doenças.
- Tabagismo e horas de sono também não se mostraram relacionados aos desfechos de saúde.

Esses achados sugerem que os hábitos analisados, isoladamente, podem não ser bons preditores dos indicadores avaliados, ao menos dentro do contexto de um dataset sintético e limitado.

Limitações

- Origem dos Dados: O dataset foi gerado por Inteligência Artificial, o que pode resultar em padrões artificiais que não refletem a realidade.
- Qualidade dos Dados: A necessidade de imputar valores e padronizar dados pode introduzir viés nos resultados.

- Causalidade: A análise exploratória não permite afirmar relações causais entre variáveis.