Relatório do desafio de Cientista de Dados

PARTE 1

Nesta primeira etapa, explorei e analisei a estrutura do dataset Dados_ficha.csv para compreender os tipos de variáveis disponíveis e identificar padrões no preenchimento dos dados. Normalmente, uma análise exploratória incluiria diversas outras investigações simultâneas, mas optei por estruturar esta resolução de forma segmentada para evidenciar meu processo de entendimento do problema. Dessa forma, cada parte do desafio reflete um aprofundamento progressivo da análise.

Passos realizados na Parte 1

1 - Carregamento e visualização inicial dos dados

Utilizei a função head() para visualizar as primeiras entradas do dataset e compreender a organização das informações. Isso permitiu identificar rapidamente que o dataset contém uma combinação de variáveis numéricas, categóricas e binárias.

2 - Identificação dos tipos de dados

Com a função dtypes, determinei os tipos das colunas (object, int64, float64) e relacionei cada tipo com sua natureza:

- Numéricas: altura, peso, pressao_sistolica, pressao_diastolica, entre outras;
- Categóricas: sexo, raca_cor, bairro, escolaridade, ocupacao etc;
- Binárias: obito, em_situacao_de_rua, luz_eletrica;
- Datas: data_cadastro, data_nascimento, updated_at;

3 - Identificação de valores ausentes

Utilizando isnull().sum(), detectei a presença de valores ausentes, principalmente em variáveis como identidade_genero, altura e peso. A análise desses valores é crucial para

compreender possíveis falhas no preenchimento e padrões de dados faltantes, preparando o terreno para inferências na Parte 2.

4 - Classificação adequada das variáveis

Algumas variáveis exigiram uma análise mais cuidadosa para determinar sua natureza. O caso do bairro, por exemplo, foi tratado como categórico, pois representa um conjunto de valores pré-definidos e não uma variável textual livre. Essa organização ajudará nas etapas seguintes ao aplicar métodos estatísticos e detectar inconsistências na base de dados.

Preparação para a Parte 2

A análise exploratória realizada nesta etapa criou um alicerce sólido para a próxima fase do desafio. A partir do entendimento da estrutura dos dados, na Parte 2 o foco será examinar mais profundamente:
Características que chamam atenção no dataset, como padrões inesperados, inconsistências e valores fora do esperado; Problemas potenciais na ingestão e coleta de dados, avaliando a confiabilidade da base e possíveis falhas no processo de preenchimento; Questionamentos relevantes aos fornecedores da informação, levantando hipóteses sobre como certos problemas podem ter sido gerados e sugerindo maneiras de melhorar a coleta de dados.

Com essa abordagem estruturada, garanto que cada etapa da análise seja conduzida de forma clara e lógica, demonstrando tanto a compreensão técnica do problema quanto a capacidade analítica necessária para interpretá-lo criticamente.

PARTE 2

1 - Características que chamam a atenção no dataset

Durante a exploração inicial do dataset, algumas características se destacaram:

- Quantidade de valores vazios: Algumas colunas possuem uma grande quantidade de valores ausentes ou listas vazias, como "doenças/condições", "meios de transporte",

"atividade física" e "tipo de atendimento", o que pode indicar que o dado não foi coletado corretamente ou que não é aplicável para certos indivíduos.

- Padrões em respostas múltiplas: Algumas variáveis categóricas permitem múltiplas respostas (como "meios de transporte" e "doenças/condições"), e essas respostas estão representadas de formas variadas (listas separadas por vírgula, ponto e vírgula, ou até espaços, além do uso de abreviações inconsistentes). Para resolver isso, seria necessário padronizar os separadores e mapear as variações de respostas para um formato único.
- Renda familiar: Existem categorias inesperadas dentro dessa variável, como "Manhã" e "Internet", sugerindo erros de digitação ou interpretação errada do campo.
- Registro de atendimentos: O número de atendimentos na atenção primária e hospitalar varia bastante, com alguns indivíduos apresentando valores significativamente altos, como registros acima de 100 atendimentos, o que pode indicar erro ou casos críticos.

2 - Problemas Identificados no Dataset

Com base nos achados, identificamos os seguintes problemas:

- Erros de digitação ou mapeamento de dados: Algumas variáveis contêm categorias que não fazem sentido, como os valores errôneos na renda familiar (exemplo: "Internet") e na situação profissional (exemplo: "N/A" em um campo que deveria ter apenas ocupações).
- Inconsistências em valores numéricos: Alturas menores que 50 cm ou maiores que 250 cm e pesos superiores a 200 kg são provavelmente erros de digitação ou medição. No entanto, alturas abaixo de 50 cm podem indicar bebês ou pessoas com condições médicas específicas. Para esclarecer isso, seria útil analisar a frequência desses valores e cruzá-los com outras variáveis, como idade e peso, para verificar se fazem sentido.
- Valores não padronizados: Algumas categorias apresentam diferentes formatos para respostas múltiplas, com caracteres especiais e variações ortográficas, como ocorre nas variáveis "meios de transporte" e "doenças/condições".
- Falta de documentação precisa: Algumas colunas possuem descrições vagas, dificultando a correta interpretação dos dados. Exemplos incluem "tipo de atendimento", "grau de instrução" e "motivo da consulta", cujos valores possíveis não

estão claramente definidos. Para resolver isso, seria necessário obter um dicionário de dados mais detalhado ou realizar inferências baseadas na distribuição das respostas.

3 - Questionamentos e Inferências

Diante dos problemas identificados, algumas perguntas e suposições podem ser levantadas:

- Sobre a coleta dos dados: Como os dados foram registrados? Houve padronização nas respostas ou foram coletados manualmente?
- Sobre inconsistências numéricas: Os valores incoerentes de altura e peso podem ter sido causados por erro humano na digitação? Existe algum mecanismo de validação para impedir esses erros?
- Sobre categorias inconsistentes: Como garantir que a renda familiar e a situação profissional sejam registradas corretamente? Existe um conjunto fechado de opções que deveriam ter sido utilizadas?
- Sobre dados faltantes: A ausência de informação é um problema na coleta ou certas informações não se aplicam a determinados indivíduos? Como lidar com esses valores vazios na análise?

4 - Possíveis Análises Complementares

Para aprimorar a qualidade da análise e identificar possíveis soluções, algumas abordagens adicionais podem ser realizadas:

- Validação de Dados: Implementar regras para detectar valores impossíveis (como alturas ou pesos irrealistas) e verificar erros de digitação nas categorias de texto.
- Tratamento de Dados Categóricos: Padronizar respostas múltiplas e revisar inconsistências na classificação de categorias.
- Imputação de Valores Ausentes: Avaliar estratégias para lidar com dados faltantes, como preenchimento baseado em distribuição ou uso de valores padrão.
- Análise de Outliers: Investigar registros que apresentam valores extremos nos números de atendimentos hospitalares e de atenção primária para entender se refletem uma realidade ou um erro.

Essa análise crítica nos permite identificar fragilidades no dataset e propor melhorias na coleta e estruturação dos dados, contribuindo para uma tomada de decisão mais assertiva.