UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS

RELATÓRIO DE TAREFA

BIOCO 01 - DADOS DA ATIVIDADE	
I - Tema Tarefa 02	**II -Turma** T01 - CIÊNCIA DOS DADOS

BIOCO 02 - DADOS DE IDENTIFICAÇÃO DOS INTEGRANTES		
III - Nome MARIA FERNANDA AQUINO FREITAS SCARCELA	**IV - Matrícula** 555889	**V - E-mail** fernandascla@alu.ufc.br
VI - Nome GIOVANNA MARIA VERISSIMO XAVIER	**VII - Matrícula** 555672	**VIII - E-mail** giovannaverissimox@alu.ufc.br
IX - Nome KLÊNISSON MATEUS OLIVEIRA RIBEIRO	**X - Matrícula** 512323	**XI - E-mail** klenissonmateuspessoal@gmail.com
XII - Nome NIVEA HAYANE GOMES MIRANDA	**XIII - Matrícula** 500028	**XIV - E-mail** niveahaiane@gmail.com

BIOCO 03 - DADOS UTILIZADOS: STROKE PREDICTION

1. **Link**: Stroke Prediction Dataset
2. **Descrição do dataset**: Esse conjunto de dados, baseado em informações da Organização Mundial da Saúde (OMS), tem como objetivo prever a probabilidade de um paciente sofrer um acidente vascular cerebral (AVC). Ele reúne dados de pacientes contendo variáveis demográficas e clínicas, como gênero, idade, presença de hipertensão e doenças cardíacas, estado civil, tipo de trabalho, tipo de residência, nível médio de glicose no sangue, índice de massa corporal (IMC) e histórico de tabagismo. Cada linha do conjunto representa um paciente e indica se ele já teve ou não um AVC, permitindo que modelos de aprendizado de máquina sejam treinados para identificar padrões e fatores de risco associados à ocorrência de AVCs.
3. **Quantidade de Colunas**: 12
4. **Quantidade de registros**: 5110
5. **Disponibilizado por: Kaggle: Your Machine Learning and Data Science Community**
6. **Licença dos Dados:** Os dados só podem ser usados para fins educacionais e o autor pede que seja citado caso o conjunto de dados seja utilizado em pesquisas ou trabalhos.
7. **Questões de Ética e Privacidade:** O dataset contém dados sensíveis de saúde, como hipertensão, doenças cardíacas, nível de glicose, IMC, tabagismo e ocorrência de AVC. Pela Lei Geral de Proteção de Dados (LGPD), esse tipo de informação exige cuidado porque pode afetar a privacidade dos indivíduos. Mesmo sem nome ou endereço, atributos como idade, gênero, estado civil e tipo de trabalho, em conjunto, podem permitir identificação indireta. Além disso, análises feitas com esses dados devem evitar gerar estigmas ou interpretações discriminatórias sobre grupos específicos. Por isso, o uso do dataset envolve cuidados éticos e riscos de privacidade, devendo ser utilizado apenas para fins educacionais ou de pesquisa, conforme indicado pelo autor.

BIOCO 04 - OBJETIVO

O objetivo deste trabalho é aplicar técnicas de Análise Exploratória de Dados (Exploratory Data Analysis – EDA) e Visualização de Dados para investigar o Stroke Prediction Dataset, que reúne informações demográficas e clínicas relacionadas ao risco de Acidente Vascular Cerebral (AVC). Por meio da exploração estatística, identificação de padrões e elaboração de visualizações informativas, busca-se compreender a distribuição das variáveis, possíveis

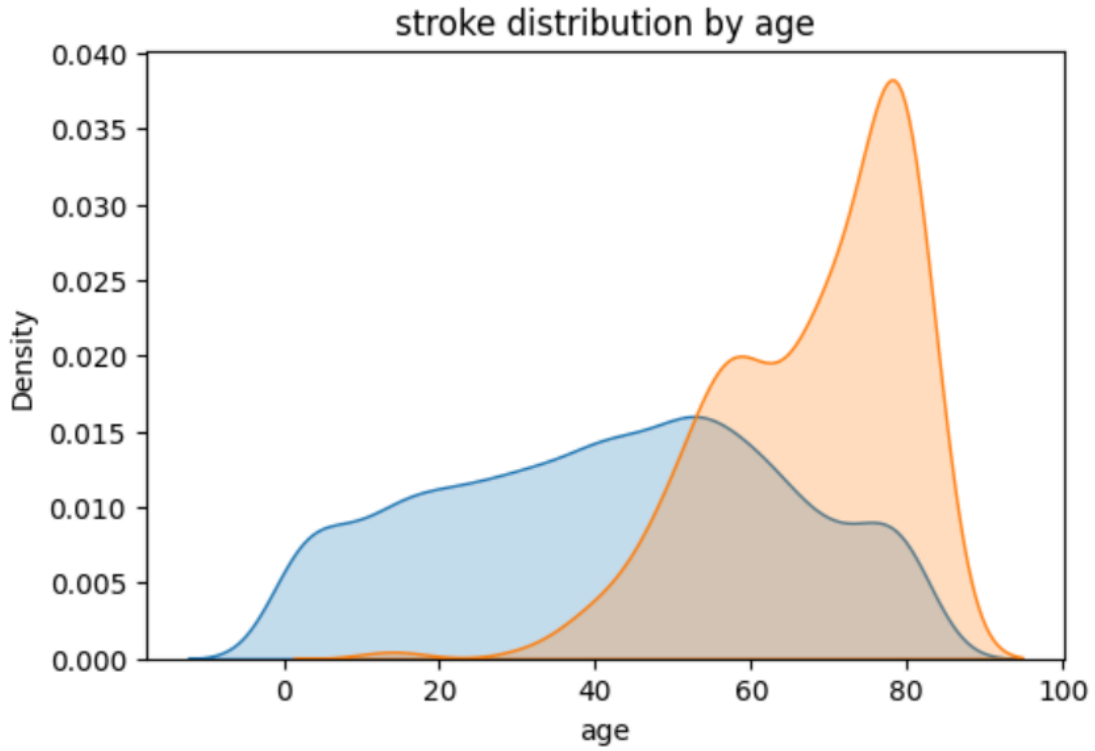
relações entre fatores de risco e a ocorrência de AVC, desenvolvendo habilidades práticas essenciais para análise e interpretação de dados de saúde.

BIOCO 05 - PROCESSAMENTOS FEITOS NOS DADOS

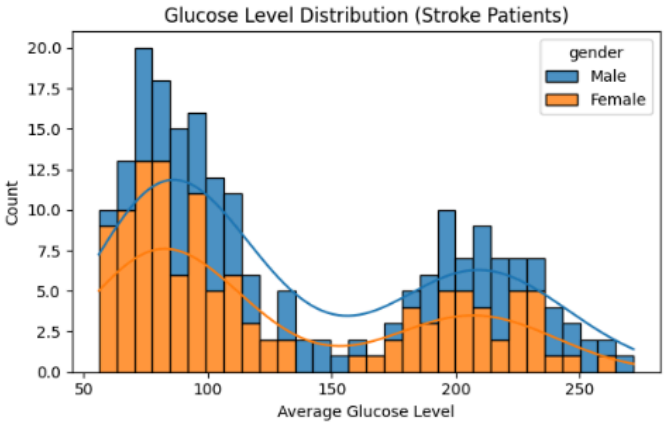
1. Inicialmente, foi realizada a verificação de valores faltantes no conjunto de dados. Identificou-se que apenas o atributo BMI apresentava valores nulos, correspondendo a aproximadamente 4% das instâncias. Dada a baixa proporção e visando manter a consistência das análises, optou-se pela remoção dessas linhas.
2. Além disso, o atributo BMI também continha valores anômalos, fora do intervalo biologicamente plausível. Esses casos foram tratados como outliers. Para corrigir as inconsistências, substituiu-se esses valores pelo valor 60, considerado um limite superior razoável associado à obesidade grave, garantindo maior coerência no conjunto de dados.
3. Por fim, durante a análise dos tipos de variáveis, observou-se que o atributo idade estava armazenado como float64. Como se trata de uma variável naturalmente inteira, realizou-se a conversão para o tipo int64, de modo a representar adequadamente a natureza dos dados.

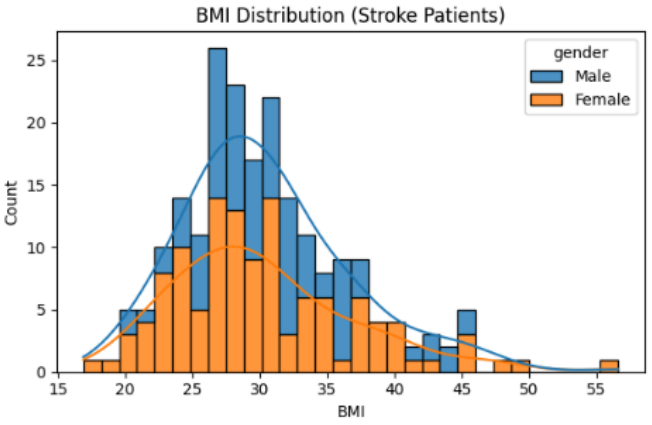
BIOCO 06 - INSIGHTS PRINCIPAIS

- Observa-se que a população sem AVC está distribuída de forma ampla ao longo das faixas etárias, desde a infância até a velhice. Já a curva dos pacientes com AVC apresenta um pico acentuado entre 60 e 80 anos, evidenciando que a probabilidade de ocorrência de AVC aumenta de maneira expressiva com o avanço da idade.

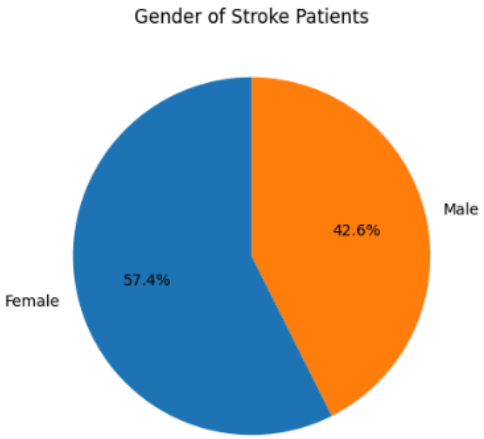


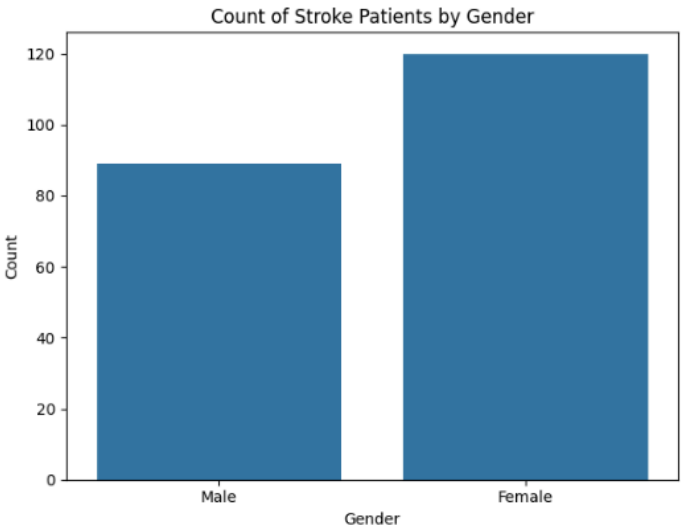
- A análise das distribuições de glicose e IMC entre pacientes com AVC nos gráficos abaixo, mostram que níveis elevados de glicose e valores de IMC na faixa de sobrepeso/obesidade são comuns nessa população. Homens tendem a apresentar valores mais altos em ambos os indicadores, sugerindo maior predominância de fatores metabólicos nos casos masculinos.



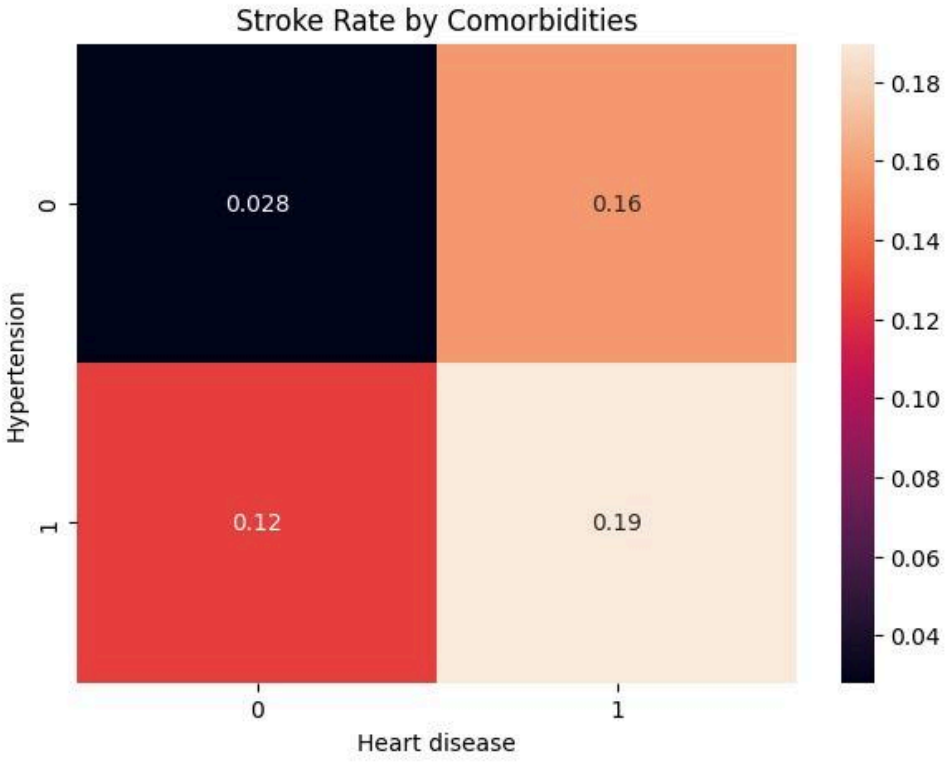


- Os dois gráficos (pizza e barras) abaixo apresentam a proporção e a contagem de pacientes com AVC separados por gênero. A análise por gênero indica que as mulheres representam a maior parte dos casos de AVC no dataset (57,4%), número superior ao observado entre os homens (42,6%).

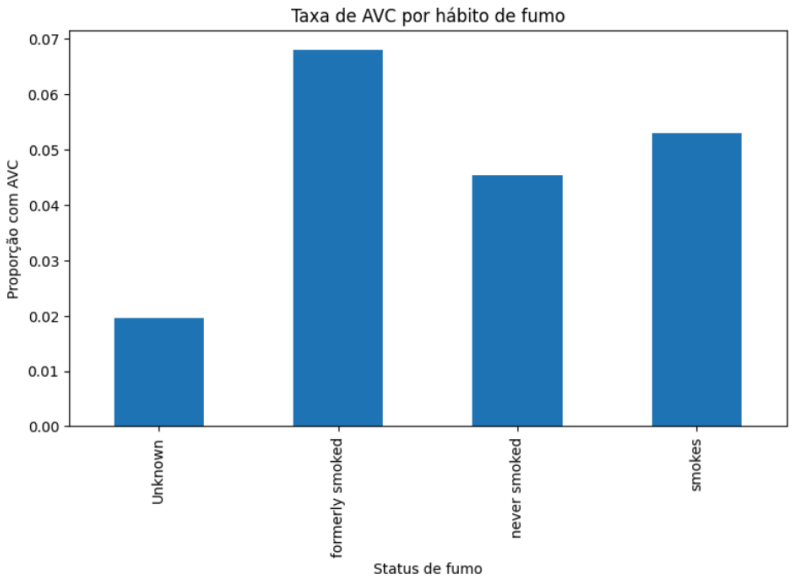




- O heatmap revela que tanto a hipertensão quanto a doença cardíaca aumentam a probabilidade de AVC, sendo a doença cardíaca o fator mais forte quando consideradas isoladamente. A combinação das duas condições resulta na maior taxa de AVC observada. Pacientes sem nenhuma dessas comorbidades apresentam risco significativamente menor.



- A análise do tabagismo mostra que esse é um dos fatores comportamentais relevantes para o risco de AVC no conjunto de dados. Observa-se que fumantes e ex-fumantes apresentam uma proporção maior de casos de AVC quando comparados a pessoas que nunca fumaram. Isso indica que tanto o hábito atual de fumar quanto o histórico de tabagismo aumentam a probabilidade de ocorrência de AVC.



BIOCO 07 - LIMITAÇÕES E RECOMENDAÇÕES

Limitações: O conjunto de dados, embora suficiente para os objetivos deste trabalho, apresenta limitações importantes. Primeiramente, dispõe de poucas variáveis clínicas, o que restringe a realização de análises mais aprofundadas sobre outros fatores potencialmente associados ao risco de AVC. Além disso, a presença de categorias pouco informativas, como o status de tabagismo classificado como "Unknown", compromete parcialmente a precisão das interpretações. Por fim, observa-se um desbalanceamento entre os grupos com e sem AVC, já que o número de casos positivos é relativamente pequeno em relação ao total de pacientes.

Recomendações: Recomenda-se, para análises futuras, complementar o estudo com outras variáveis clínicas mais específicas, caso estejam disponíveis, e aplicar métodos de balanceamento para lidar melhor com a diferença entre as classes. Também seria útil trabalhar com mais conjuntos de dados ou comparar os resultados com outras bases para reforçar as conclusões.

BIOCO 08 - DESAFIOS E APRENDIZAGEM

Desafios: Um dos desafios do trabalho foi criar perguntas e hipóteses relevantes sobre os dados, conectando cada uma delas a possíveis benefícios sociais ou decisões práticas que poderiam ser tomadas a partir dos insights. Outro desafio importante foi escolher visualizações adequadas para responder cada pergunta. Encontrar o gráfico certo para destacar padrões como a relação entre idade, comorbidades e risco de AVC demandou tentativas, ajustes e compreensão das melhores formas de representar cada tipo de informação.

Aprendizagem: Como aprendizados, o projeto mostrou a importância da limpeza dos dados, da escolha adequada dos gráficos e de analisar cada variável com cuidado para evitar conclusões erradas. Também ajudou a entender melhor como a EDA revela padrões importantes para a compreensão do risco de AVC.