Previsão de AVC usando Regressão Logística: Um Estudo de Características Clínicas.

Larissa Vitória Vasconcelos Sousa
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
larissavvsousa@alu.ufc.br

Mateus Silva Matos
Itapajé, CE-Brazil
mateussilvamatos@alu.ufc.br

***Abstract*—This article presents an analysis of two datasets related to stroke and diabetes. The objective is to explore and classify the raw data, identify common behaviors, and provide insights into the characteristics and patterns observed. The preprocessing of the data involves steps to improve data quality, such as handling missing values, translating categorical variables, and creating age groups. The exploration and visualization of the data reveal associations and correlations between variables, particularly in relation to age and glucose levels. The results show that the datasets exhibit similarities and differences in terms of age distribution, presence of risk factors, and prevalence of the conditions. These findings contribute to a better understanding of the health conditions and risk factors associated with stroke and diabetes. The article emphasizes the importance of data preprocessing in ensuring data quality and validity for further analysis and modeling. The insights gained from this study can support the development of preventive strategies, early diagnosis, and effective treatment for these health conditions.**

***Index Terms*—Stroke, preprocessing, exploration, visualization.**

I. Introdução

Segundo a Organização Mundial da Saúde (OMS), o AVC é a 2ª principal causa de morte no mundo, responsável por aproximadamente 11% do total de mortes. O dataset "Stroke Prediction Dataset" apresenta um conjunto de dados disponível no Kaggle é usado para prever se um paciente tem probabilidade de sofrer um derrame com base nos parâmetros de entrada, como sexo, idade, várias doenças e tabagismo. O dataset contém informações demográficas e clínicas de indivíduos, incluindo idade, sexo, hipertensão, doença cardíaca, tabagismo, tipo de trabalho, estado civil, nível de glicose, índice de massa corporal (IMC) e status de atividade física. Além disso, é fornecido o registro sobre se um indivíduo teve ou não um AVC.

A análise de pré-processamento do banco de dados brutos sobre AVC é necessária devido à importância de garantir a qualidade e a confiabilidade dos dados utilizados para análises e modelagem. O pré-processamento desempenha um papel fundamental na identificação e tratamento de problemas como valores ausentes, dados inconsistentes, variáveis categóricas não tratadas e desequilíbrio de classes. Esses problemas podem comprometer a precisão e a validade dos resultados obtidos a partir dos dados brutos.

Trabalho Avaliativo da disciplina de Laboratório de Ciência de Dados

O pré-processamento de dados é uma etapa essencial em qualquer análise de dados ou modelagem estatística. Pesquisas científicas e estudos acadêmicos em diferentes áreas têm enfatizado a importância do pré-processamento de dados para garantir a qualidade e a validade dos resultados. Métodos e técnicas de pré-processamento têm sido amplamente discutidos e utilizados em diversos campos, como medicina, economia, ciência da computação e ciências sociais.

II. Pré-processamento dos dados

A. Descrição da mineração dos dados

O pré-processamento dos dados sobre AVC envolve várias etapas para melhorar a qualidade e usabilidade das informações. Inicialmente, os nomes das colunas são traduzidos para o português, facilitando a compreensão dos dados. Em seguida, os valores categóricos são traduzidos para termos mais compreensíveis, como gênero (de "Male" e "Female" para "Masculino" e "Feminino"). Idades menores que 1 são analisadas e conclui-se que são bebês ou crianças, assim, a condição de fumante é atribuída como "Desconhecido" para elas. Essa abordagem aborda inconsistências relacionadas a idades inconsistentes. A criação de faixas etárias e categorias para o nível médio de glicose permite uma análise mais fácil. Valores ausentes (NaN) são identificados e tratados, permitindo a análise adequada dos dados. Em suma, o pré-processamento tem como objetivo melhorar a qualidade e consistência dos dados para análises futuras e modelagem final.

B. Exploração e Visualização dos dados

Nesta seção, será realizada a exploração e visualização dos dados. Serão apresentados gráficos que ajudarão na realização de uma análise descritiva e exploratória dos dados. Em seguida, um modelo de regressão logística será treinado utilizando o conjunto de dados fornecido, com o objetivo de prever se uma pessoa terá um AVC (Acidente Vascular Cerebral) ou não, levando em consideração características como idade, sexo, pressão arterial, nível de glicose, entre outras. Essa análise proporcionará insights valiosos sobre os fatores

de risco associados ao AVC e permitirá o desenvolvimento de um modelo preditivo para auxiliar na identificação precoce de indivíduos com maior probabilidade de desenvolver essa condição.

A seguir, a tabela 1 exibe o número de pacientes - após todo o pré processamento - categorizados de acordo com faixas etárias estabelecidas.

TABLE I
NÚMERO DE PACIENTES POR FAIXA ETÁRIA

Faixas etárias	Número de pacientes
Crianças (0 a 9 anos)	472
Jovens (10 a 19 anos)	494
Adultos (20 a 49 anos)	1.934
Idosos (acima de 50 anos)	2.210
Total	5.110

A primeira coluna indica as faixas etárias, incluindo crianças (0 a 9 anos), jovens (10 a 19 anos), adultos (20 a 49 anos) e idosos (acima de 50 anos). A segunda coluna mostra o número de pacientes encontrados em cada faixa etária. Por exemplo, há 472 crianças, 494 jovens, 1.934 adultos e 2.210 idosos. O total de pacientes é igual a 5.110.

Será apresentada a análise descritiva por meio de tabelas, fornecendo um panorama detalhado das características e tendências encontradas na amostra de dados. Cada tabela oferece informações relevantes para compreender os padrões e insights do conjunto de dados. As tabelas são uma ferramenta poderosa de visualização de dados, permitindo identificar padrões, tendências e relações importantes. Agora, exploraremos cada tabela e realizaremos a análise descritiva correspondente.

A Tabela 2 apresenta a distribuição do tipo de trabalho dos pacientes em diferentes faixas etárias. O tipo de trabalho pode variar entre crianças, trabalho privado, nunca ter trabalhado, trabalho autônomo e trabalho no governo.

TABLE II
TIPO DE TRABALHO DOS PACIENTES

Faixas etárias	Variável				
	Tipo de trabalho				
	Criança	Privado	Nunca	Autônomo	Governo
Crianças	99,58%	0,21%	0	0,21%	0
Jovens	43,93%	46,56%	4,25%	3,24%	2,02%
Adultos	0	75,08%	0,05%	10,08%	0
Idosos	0	56,20%	0	27,47%	16,33%

Entre as crianças, é esperado que a maioria - ou todas - não estejam envolvidas em trabalho, entretanto uma pequena proporção está envolvida em trabalho privado ou autônomo. Nos jovens, há uma distribuição mais equilibrada entre trabalho privado, nunca ter trabalhado, trabalho autônomo e trabalho no governo. Nos adultos, a maioria está envolvida em trabalho privado, seguido por trabalho autônomo, e uma pequena porcentagem nunca trabalhou. Entre os idosos, observa-se uma participação significativa no trabalho privado e trabalho no governo, com uma parcela considerável envolvida em trabalho autônomo.

A seguir, é apresentada a distribuição do tipo de residência dos pacientes em diferentes faixas etárias, essa análise permitirá compreender como o ambiente residencial está distribuído entre os grupos.

TABLE III
TIPO DE RESIDÊNCIA DOS PACIENTES

Faixas etárias	Variável	
	Tipo de residência	
	Rural	Urbano
Crianças	47,67%	52,33%
Jovens	53,04%	46,96%
Adultos	49,79%	50,21%
Idosos	48,14%	51,86%

Na tabela acima, observa-se que entre as crianças, a maioria reside em áreas urbanas, representando 52,33% do total, enquanto 47,67% vivem em áreas rurais. Já entre os jovens, há uma distribuição relativamente equilibrada, com 46,96% residindo em áreas urbanas e 53,04% em áreas rurais. Ao analisar os adultos, nota-se uma proporção bastante semelhante entre residências rurais (49,79%) e urbanas (50,21%). Quanto aos idosos, a tendência se mantém, com 51,86% vivendo em áreas urbanas e 48,14% em áreas rurais.

A seguir, a tabela 4 mostra a distribuição do tabagismo entre pacientes, divididos em diferentes faixas etárias. A variável analisada é a condição fumante, que engloba quatro categorias distintas: desconhecida, nunca fumou, fumava antes e fuma atualmente.

TABLE IV
TABAGISMO EM PACIENTES

Faixas idades	Variável			
	Condição fumante			
	Desconhecida	Nunca fumou	Fumava antes	Fuma
Crianças	100%	0	0	0
Jovens	52,43%	37,85%	5,87%	3,85%
Adultos	20,79%	43,49%	14,27%	21,46%
Idosos	18,60%	39,10%	26,24%	16,06%

Cada célula da tabela representa a proporção de pacientes dentro de uma faixa etária e categoria de condição fumante. Observa-se que as crianças não possuem nenhum paciente fumante, enquanto os jovens apresentam uma distribuição mais equilibrada entre as categorias. Os adultos têm uma proporção considerável de pacientes fumantes, principalmente na categoria de ”fuma atualmente”. Já os idosos apresentam uma distribuição mais diversificada, com um número significativo de pacientes que fumavam antes. Essa tabela permite uma análise detalhada para a compreensão desse comportamento em relação à saúde dos pacientes.

Na tabela a seguir será apresentado a distribuição do estado civil entre os pacientes, segmentados por faixas etárias. A variável analisada é o estado civil, especificamente se os pacientes são casados ou não.

Todas as crianças da amostra não são casadas, representando 100% da faixa etária. Nos jovens, a maioria também não é casada, com apenas 0,61% dos pacientes nessa condição.

TABLE V
ESTADO CIVIL DOS PACIENTES

Faixas etárias	Variáveis	
	Estado civil (casado)	
	Sim	Não
Crianças	0	100%
Jovens	0,61%	99,39%
Adultos	67,84%	32,16%
Idosos	92,22%	7,78%

Nos adultos, a proporção de pacientes casados aumenta significativamente, chegando a 67,84%. Entre os idosos, a grande maioria, 92,22% dos pacientes, é casada.

Seguidamente, será apresentada a análise da existência de AVC (Acidente Vascular Cerebral) entre os pacientes, divididos em diferentes faixas etárias.

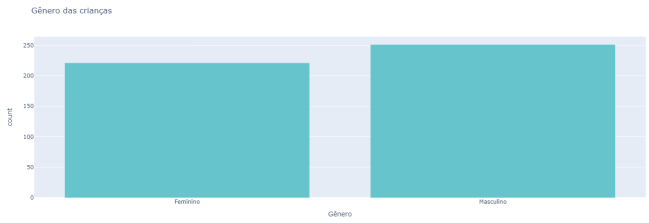
TABLE VI
EXISTÊNCIA DE AVC NOS PACIENTES

Faixas etárias	Variável	
	AVC	
	Sim	Não
Crianças	0,21%	99,79%
Jovens	0,20%	99,80%
Adultos	0,93%	99,07%
Idosos	10,36%	89,64%

Observa-se que as crianças e os jovens têm uma baixa proporção de casos de AVC, com valores próximos a 0,2% em ambas as faixas etárias. Nos adultos, a proporção de pacientes com AVC é ligeiramente maior, atingindo 0,93%, enquanto nos idosos, o percentual de casos de AVC é mais significativo, chegando a 10,36%. Esses dados sugerem que a ocorrência de AVC está mais presente entre os pacientes idosos, tornando-se um aspecto relevante para estratégias preventivas e intervenções médicas nessa faixa etária. Essa análise descritiva permite uma melhor compreensão da distribuição do AVC em relação à idade dos pacientes e fornece informações cruciais para o aprimoramento do cuidado e tratamento desses indivíduos.

Na Figura 1, será realizada uma análise do gênero dos pacientes na faixa etária infantil.

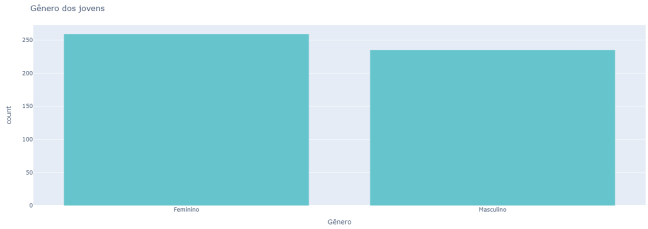
Fig. 1. Gênero: Crianças



No grupo das crianças entre 0 e 9 anos, observa-se uma leve predominância de meninos em relação às meninas.

Na Figura 2, será realizada uma análise do gênero dos pacientes na faixa etária adolescente/jovens.

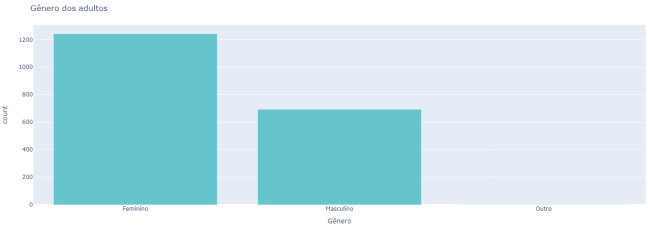
Fig. 2. Gênero: Jovens



No grupo dos adolescentes/jovens entre 10 e 19 anos, é possível notar uma ligeira predominância de meninas em relação aos meninos.

Na Figura 3, será realizada uma análise do gênero dos pacientes na faixa etária adultos.

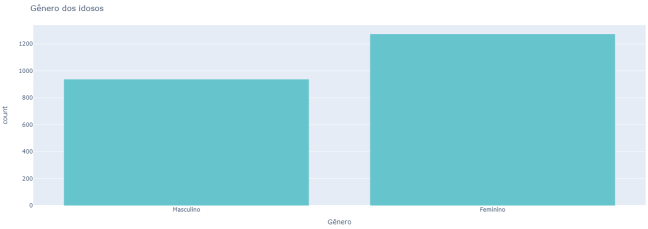
Fig. 3. Gênero: Adultos



No grupo de adultos, compreendendo indivíduos com idades entre 20 e 49 anos, é notável a presença de mais de 1000 mulheres registradas, em comparação com o número de homens que alcança 700. É relevante destacar que apenas 1% dos pacientes identificam-se com o gênero "Outro". Essas informações ressaltam a disparidade de gênero nessa faixa etária e a predominância feminina no conjunto de dados.

Na Figura 4, uma análise será realizada do gênero dos pacientes da faixa etária idosos.

Fig. 4. Gênero: Idosos



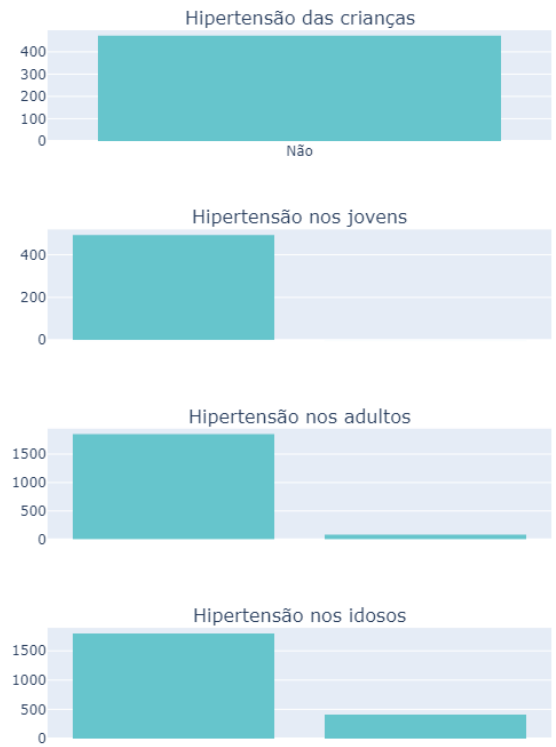
No grupo de idosos, composto por pacientes com idades acima de 50 anos, destaca-se a presença de mais de 1200 mulheres registradas, enquanto os homens, totalizam mais de 900.

Fica evidente que a predominância é do gênero feminino entre os pacientes. No entanto, uma exceção ocorre no grupo das crianças, onde há uma proporção ligeiramente maior de meninos em relação às meninas. Já no grupo de adultos, é interessante notar que existem mais de 1000 mulheres registradas, enquanto o número de homens chegam a 500. Além

disso, é importante mencionar que apenas 1% dos pacientes se identifica com o gênero ”Outro”.

Na Figura 5, será apresentado a presença ou ausência de hipertensão nos pacientes.

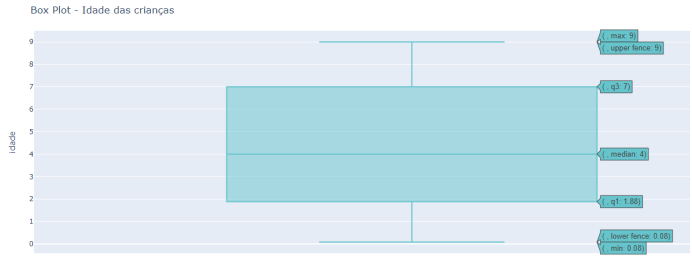
Fig. 5. Hipertensão por faixa etária



As crianças não tem hipertensão (Não: 100%), adultos e idosos já possuem uma porcentagem significativa de que têm hipertensão, idosos são quase 500 pacientes.

Na Figura 6, serão apresentados Box Plots de idade em cada faixa etária.

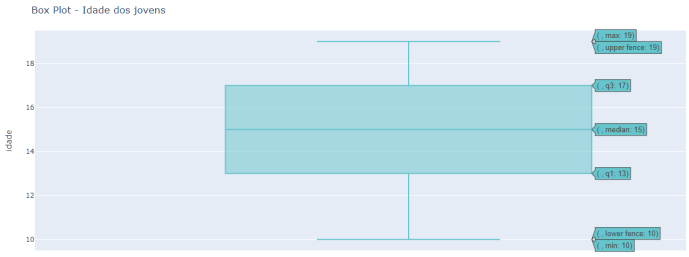
Fig. 6. Idades das crianças



Os valores estatísticos incluem o máximo (9), que representa o maior valor observado; o terceiro quartil (7), que indica o valor abaixo do qual 75% dos dados estão concentrados; a mediana (4), que divide a distribuição ao meio; o primeiro quartil (1.88), que delimita o valor abaixo do qual 25% dos dados estão concentrados; e o mínimo (0.08), que representa o valor mais baixo observado.

Agora será realizada uma análise do Box Plot da idade dos jovens. Essa visualização estatística proporcionará informações relevantes sobre medidas de tendência central, variação e possíveis valores atípicos, permitindo uma compreensão mais aprofundada da distribuição da idade dos jovens.

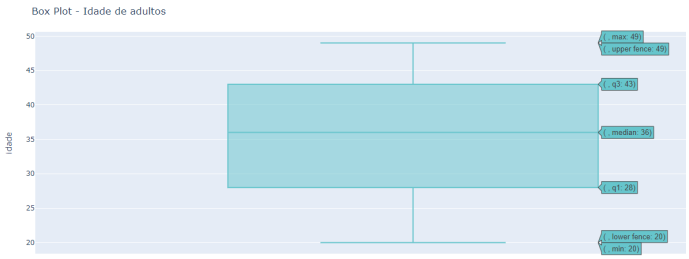
Fig. 7. Idades dos jovens



No BoxPlot da idade dos jovens, o valor máximo (max) é de 19 anos, indicando a idade mais alta observada. O terceiro quartil (Q3) é de 17 anos, representando o valor abaixo do qual 75% dos dados estão concentrados. A mediana, localizada no centro da caixa, possui o valor de 15 anos, dividindo a distribuição ao meio. O primeiro quartil (Q1) é 13, determinando o valor abaixo do qual 25% dos dados estão concentrados, e o valor mínimo (min) observado é 10 anos.

Na Figura 8, será realizada uma análise do Box Plot da idade dos adultos.

Fig. 8. Idades dos adultos

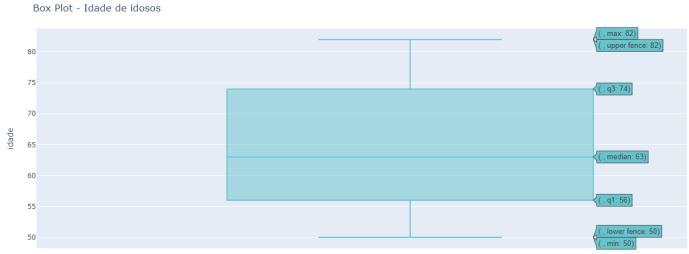


No gráfico da idade dos adultos, observa-se os seguintes valores estatísticos: o valor máximo (max) é de 49 anos, representando a idade mais alta registrada no conjunto de dados. O terceiro quartil (Q3) é de 43 anos, indicando que 75% dos adultos têm idade inferior a esse valor. A mediana está localizada em 36 anos, dividindo a distribuição ao meio. O primeiro quartil (Q1) é 28, indicando que 25% dos adultos têm idade abaixo desse valor. O valor mínimo (min) observado é 20 anos.

Na Figura 9, iremos realizar uma análise do Box Plot que representa a distribuição da idade dos idosos.

No caso dos idosos, o valor máximo é de 82 anos, indicando a idade mais avançada registrada no conjunto de dados. O terceiro quartil (Q3) é de 74 anos, indicando que 75%

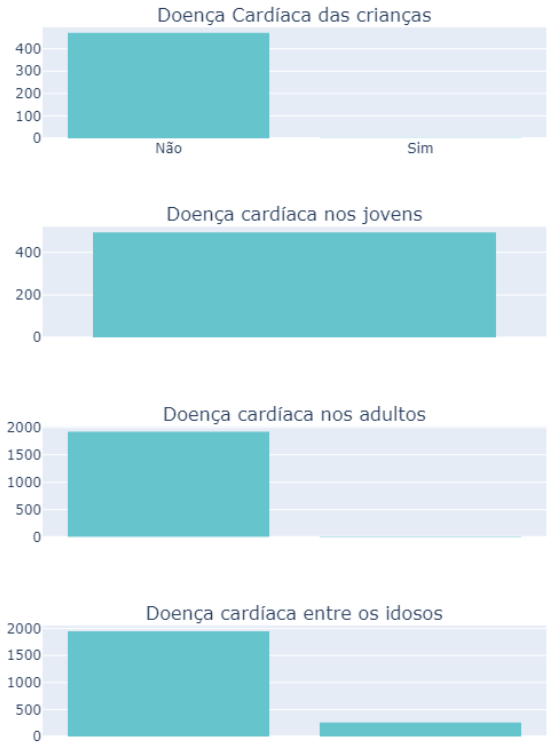
Fig. 9. Idades dos idosos



dos idosos têm idade inferior a esse valor. A mediana está localizada em 63 anos, dividindo a distribuição ao meio. O primeiro quartil (Q1) é 56, indicando que 25% dos idosos têm idade abaixo desse valor. O valor mínimo (min) observado é 50 anos.

Na Figura 10 a seguir, uma análise dos gráficos sobre os pacientes que possuem doença cardíaca será realizada.

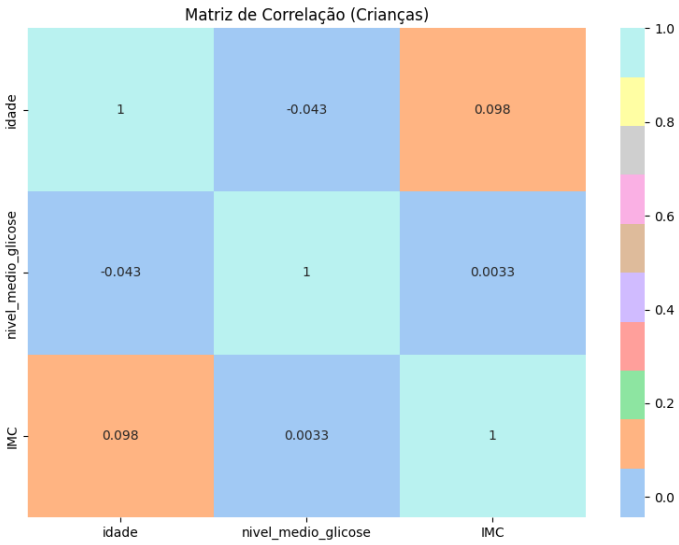
Fig. 10. Doença cardíaca por faixa etária



Ao analisar a presença de doença cardíaca em diferentes faixas etárias, percebe-se que a maioria dos pacientes não apresenta a doença. No entanto, é importante destacar que na faixa dos idosos, há uma proporção significativa de pacientes afetados por essa condição. Isso sugere que a incidência de doença cardíaca aumenta à medida que a idade avança, tornando-se mais prevalente entre os indivíduos mais velhos.

A Figura 11 mostra as relações entre idade, nível médio de glicose e IMC na faixa etária infantil.

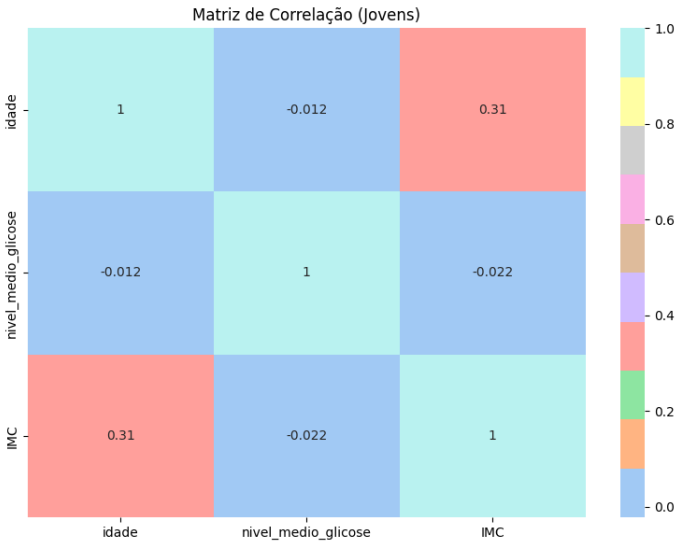
Fig. 11.



A matriz de correlação revela que, na faixa etária das crianças, há uma correlação negativa muito fraca entre o nível médio de glicose e a idade (-0.043). Da mesma forma, a correlação positiva entre o IMC e a idade é fraca (0.098), não demonstrando uma relação linear forte entre essas duas variáveis. Além disso, a correlação entre o IMC e o nível médio de glicose é quase nula (0.0033). Em suma, a matriz de correlação sugere que não há relações lineares significativas entre o nível médio de glicose, o IMC e a idade nas crianças.

A Figura 12 apresenta as relações entre as mesmas três variáveis, na faixa etária de jovens.

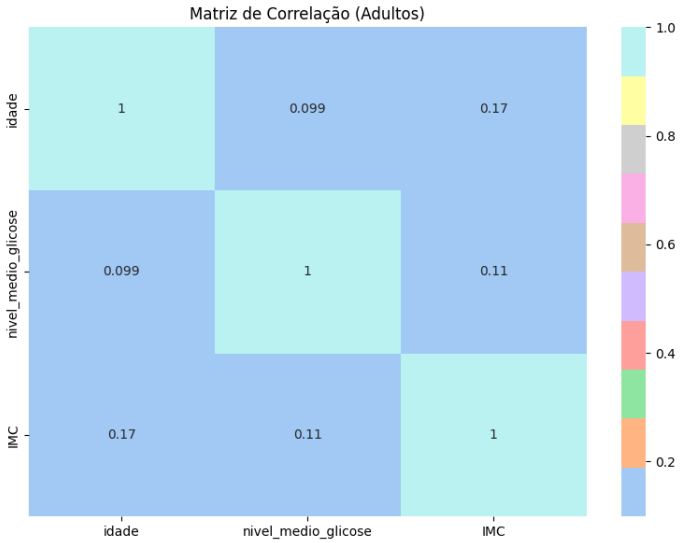
Fig. 12.



A relação entre a glicose média e a idade é de -0,012 (cor azul), indicando uma correlação negativa muito fraca. Por outro lado, a correlação entre IMC e idade é de 0,31 (cor rosa), indicando uma correlação moderadamente positiva. Isso implica uma relação linear relativamente forte entre IMC e idade em jovens, sugerindo que o IMC tende a aumentar com a idade nessa faixa etária. A correlação entre o IMC e a glicemia média é de -0,022 (cor azul), indicando uma correlação negativa muito fraca. Isso sugere que não há relação linear significativa entre o IMC e os níveis médios de glicose no sangue em jovens.

Na Figura 13 a seguir, é possível visualizar as relações entre as três variáveis: idade, nível médio de glicose e IMC (Índice de Massa Corporal) na faixa etária de adultos.

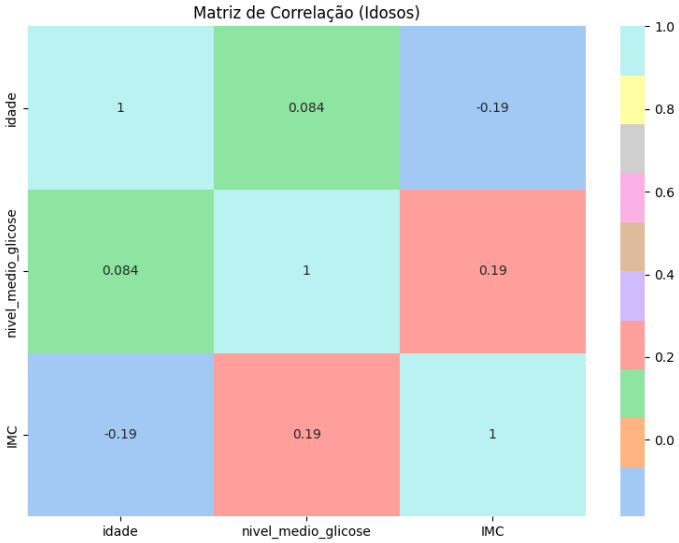
Fig. 13.



Na matriz de correlação dos adultos, a relação entre glicose média e idade foi de 0,099, indicando uma relação positiva fraca. Isso sugere que existe uma pequena relação entre os níveis médios de glicose em adultos e que conforme a idade aumenta, a glicose aumenta juntamente com a idade nessa faixa etária. Para IMC e idade, a correlação foi de 0,17, indicando uma correlação moderadamente positiva. Isso significa que existe uma relação relativamente forte entre o IMC e a idade em adultos, sugerindo que o IMC aumenta com a idade. A correlação entre IMC e glicose média foi de 0,11, indicando uma relação positiva fraca. Isso sugere uma leve associação entre o IMC e os níveis médios de glicose em adultos.

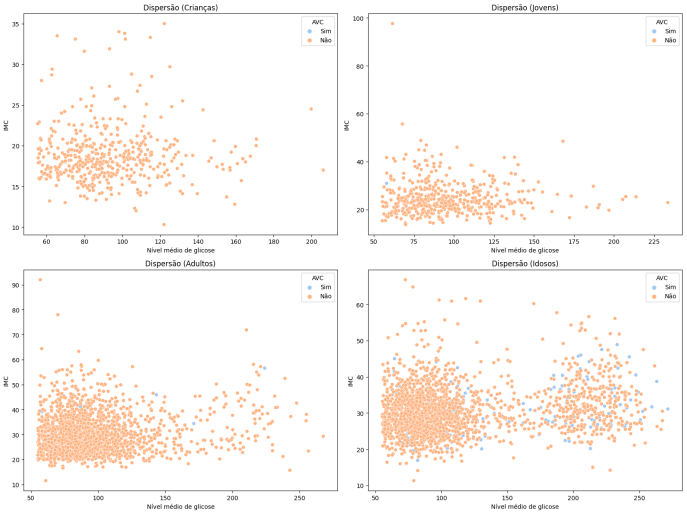
Na Figura 14, apresenta-se a representação gráfica das interações entre as variáveis idade, nível médio de glicose e IMC (Índice de Massa Corporal) na faixa etária dos idosos.

Fig. 14.



A correlação entre o nível médio de glicose no sangue e a idade é de 0,084 (cor verde), indicando uma fraca correlação positiva. Isso sugere uma pequena associação entre os níveis médios de glicose no sangue de idosos e a idade, sugerindo que os níveis de glicose no sangue nessa faixa etária tendem a aumentar ligeiramente com a idade. A correlação entre IMC e idade é de -0,19 (cor azul), indicando uma correlação moderadamente negativa. Isso significa que existe uma associação relativamente forte entre IMC e idade em idosos, sugerindo que o IMC tende a diminuir com a idade. A correlação entre o IMC e a glicemia média foi de 0,19 (cor rosa), indicando uma correlação moderadamente positiva. Isso sugere uma associação modesta entre o IMC e os níveis médios de glicose no sangue em idosos.

Fig. 15. Relação entre glicose e AVC em faixas etárias



Os gráficos acima, retratam a relação entre glicose e AVC, os pontos em azul mostram os pacientes que possuem

diabetes e AVC, é importante observar a grande quantidade na faixa etária de idosos.

MODELO: REGRESSÃO LOGÍSTICA

Após todas as análises, um modelo de Regressão Logística foi treinado com o objetivo de prever se uma pessoa terá um AVC ou não, com base em características como idade, sexo, hipertensão, nível de glicose, doença cardíaca, IMC, entre outras. A base de dados bruta é lida novamente em outro arquivo, o pré-processamento dos dados é realizado para lidar com valores ausentes e codificar variáveis categóricas, garantindo a adequação dos dados para a modelagem. Para evitar qualquer viés decorrente do desequilíbrio entre as classes, é aplicado o oversampling (técnica utilizada no campo de aprendizado de máquina e mineração de dados para lidar com conjuntos de dados desbalanceados) na classe minoritária.

O conjunto de dados é então dividido em conjuntos de treinamento e teste para avaliar o desempenho do modelo. Além disso, as variáveis numéricas são normalizadas/escaladas utilizando o StandardScaler. O modelo de Regressão Logística é definido, treinado com os dados de treinamento e posteriormente utilizado para fazer previsões nos dados de teste. Para avaliar a performance do modelo, métricas como acurácia, precisão, recall e F1-score são calculadas e exibidas.

TABLE VII
MÉTRICAS PARA AVALIAÇÃO DO MODELO

Modelo: Regressão Logística		
Métricas	**Resultados**	**Em %**
Acurácia	0.7994858611825193	79,94%
Precisão	0.7685185185185185	76,85%
Recall	0.8556701030927835	85,56%
F1-score	0.8097560975609756	80,97%

As métricas de desempenho indicam que o modelo de regressão logística possui uma acurácia de 0.799, classificando corretamente cerca de 79,9% dos casos. A precisão é de 0.769, o que significa que cerca de 76,9% das instâncias classificadas como positivas pelo modelo são realmente positivas. O recall é de 0.856, indicando que o modelo identifica corretamente cerca de 85,6% das instâncias positivas. O F1-score é de 0.810, uma medida equilibrada entre precisão e recall. Essas métricas fornecem uma avaliação do desempenho geral e específico do modelo na classificação de casos de AVC.

Por fim, é gerada e apresentada a matriz de confusão, fornecendo uma visão detalhada dos resultados obtidos pelo modelo.

TABLE VIII
MATRIZ DE CONFUSÃO

		Real	
		Tem AVC	**Não tem AVC**
Previsto	**Tem AVC**	Verd. Positivos 830	Falsos Positivos 250
	Não tem AVC	Falsos Negativos 140	Verd. Negativos 725

A matriz de confusão é uma tabela que resume as classificações feitas por um modelo de classificação para cada classe. Ela mostra o número de verdadeiros positivos, falsos positivos, falsos negativos e verdadeiros negativos. No caso apresentado, o modelo classificou corretamente 830 casos como "Tem AVC" (verdadeiros positivos), classificou incorretamente 250 casos como "Tem AVC" quando na verdade não tinham (falsos positivos), deixou de identificar corretamente 140 casos como "Tem AVC" (falsos negativos) e classificou corretamente 725 casos como "Não tem AVC" (verdadeiros negativos).

Com base nas métricas de desempenho apresentadas (acurácia, precisão, recall e F1-score) e na matriz de confusão, pode-se concluir que o modelo de regressão logística obteve um desempenho satisfatório.

C. Principais técnicas usadas na primeira etapa da disciplina: pré-processamento

- Foi utilizado Estatística Descritiva, com a função "describe" do Python, para analisar as variáveis quantitativas discretas e contínuas como contagem, média, desvio padrão, quartis, valor mínimo e máximo usando a biblioteca "Pandas";
- Avaliamos a correlação entre as variáveis usando Inferência e a biblioteca "Seaborn" para verificar as possíveis ligações entre as colunas do dataset;
- Para a visualização do comportamento dos dados, foi usada a linguagem de programação Python utilizando as bibliotecas "Matplotlib" e "Plotly".

III. AVALIANDO UM SEGUNDO *raw dataset*

O segundo dataset está relacionado a Diabetes. A influência de determinadas variaveis na possibilidade de possuir ou não a doença.uma definição mais precisa sobre diabetes: É uma doença crônica caracterizada pelo aumento dos níveis de glicose (açúcar) no sangue. Isso ocorre devido a problemas na produção ou na utilização da insulina, um hormônio secretado pelo pâncreas que regula a quantidade de glicose no corpo. A insulina permite que as células absorvam a glicose do sangue e a utilizem como fonte de energia.O estudo não foi tão aprofundado nesse dataset comparado com dataset sobre AVC, pois o estudo principal era sobre o que motivava o acontecimento do AVC na sociedade.

Fig. 16. Base de dados

	Gravidez	glicose	BloodPressure	pele	insulina	IMC	Predicao_diabetes	idade	Diabetes
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

A base de dados possui 9 variaveis, sendo as mais utilizadas ”glicose” e ”possui ou não diabetes”.

Fig. 17. Análise descritiva

	Gravidez	glicose	BloodPressure	pele	insulina	IMC	Predicao_diabetes	idade	Diabetes
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Na Fig. acima podemos ver os valores de media, mediana e quartis.

Fig. 18. Correlação das variáveis

	Gravidez	glicose	BloodPressure	pele	insulina	IMC	Predicao_diabetes	idade
Gravidez	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
glicose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
pele	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
insulina	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
IMC	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
Predicao_diabetes	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
idade	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00

Fig. 19. Idade por faixa etaria

	idade	faixa_etaria
0	50	50 a 59 anos
1	31	30 a 39 anos
2	32	30 a 39 anos
3	21	20 a 29 anos
4	33	30 a 39 anos
...	...	...
763	63	60 a 69 anos
764	27	20 a 29 anos
765	30	30 a 39 anos
766	47	40 a 49 anos
767	23	20 a 29 anos

768 rows × 2 columns

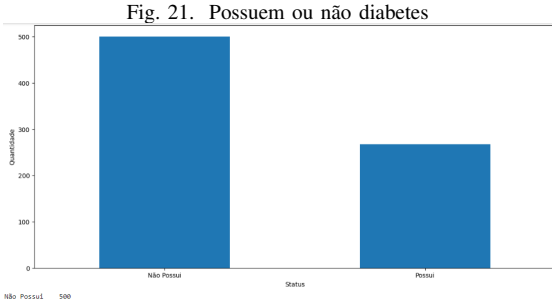
Fig. 20. Porcentagem dos que possuem ou não diabetes

```
quantidades de quem possui ou não diabetes

DIABETES

Não Possui    65.10%
Possui        34.90%
Name: Diabetes, dtype: object
```

É notável que a maior parte das pessoas estudadas não possuem diabetes.

A. *Exploração e Visualização dos dados*

Fig. 21. Possuem ou não diabetes



É notável que quanto maior for a glicose muito provavelmente essa pessoa possuirá diabetes.

Fig. 22. Usando gráfico de pizza

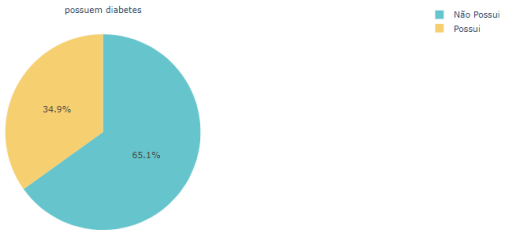


Fig. 23. Relação glicose vs diabetes

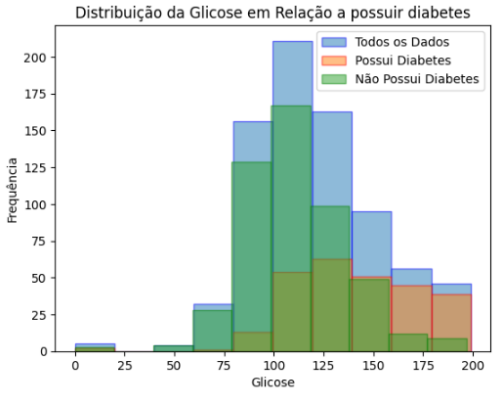


Fig. 24. Glicose

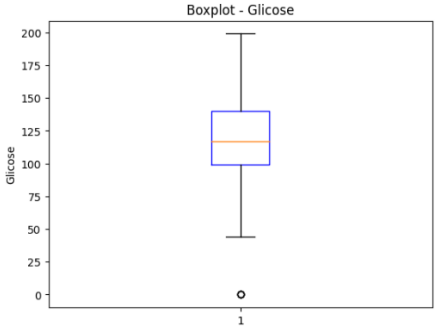


Fig. 25. Diabetes

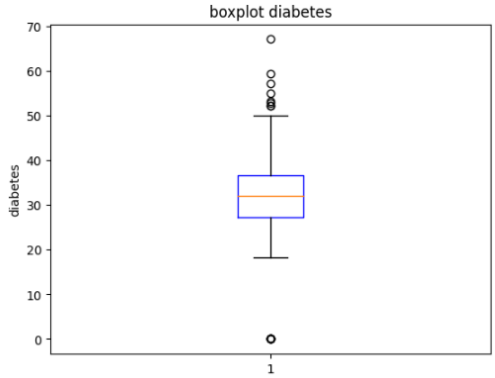


Fig. 26. Correlação de pearson

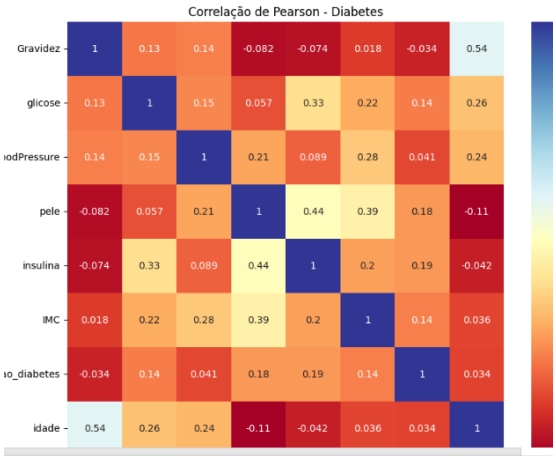


Fig. 27. Relação idade vs glicose

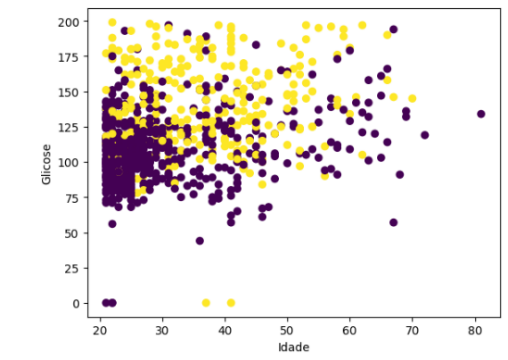
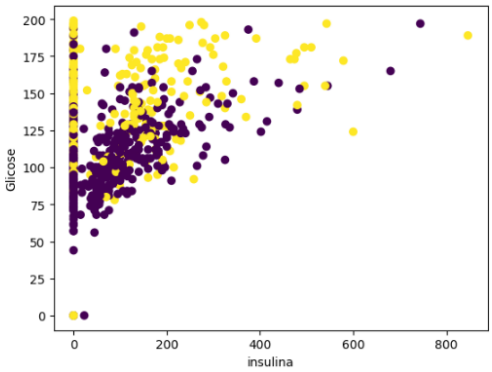


Fig. 28. Relação insulina vs glicose



IV. Resultados

Ao comparar os dois datasets, observou-se que ambos apresentaram uma distribuição variando de crianças a idosos em relação à faixa etária. No entanto, as proporções de pacientes em cada faixa etária foram diferentes. No dataset de AVC, houve uma concentração maior de pacientes idosos, enquanto no dataset de Diabetes, a distribuição foi mais equilibrada entre as faixas etárias.

Em relação às variáveis específicas de cada dataset, a glicose foi uma variável comum em ambos. Foi identificada uma correlação positiva entre os níveis de glicose e a presença de AVC, indicando um maior risco de AVC em pacientes com níveis mais altos de glicose. No dataset de Diabetes, também observou-se uma relação semelhante, sugerindo que níveis elevados de glicose estão associados à presença da doença.

Uma diferença notável entre os datasets foi a presença de informações sobre tabagismo apenas no dataset de AVC. Essa diferença limitou a análise dos fatores de risco relacionados ao tabagismo no dataset de Diabetes.

De forma geral, ambos os datasets forneceram informações valiosas sobre as características das amostras e os padrões presentes. A análise exploratória dos dados revelou associações entre variáveis e forneceu insights sobre os fatores de risco e as condições de saúde em estudo. A visualização dos dados por meio de gráficos e tabelas facilitou a compreensão e comunicação dos resultados, auxiliando na tomada de decisões informadas por profissionais de saúde.

Conclusões

Ao analisar os dados brutos dos dois datasets, encontramos comportamentos comuns que fornecem informações valiosas sobre as características das amostras e os padrões presentes. No dataset de AVC, observamos que crianças não têm casos de AVC, enquanto idosos apresentam uma proporção significativa dessa condição. Além disso, notamos uma proporção considerável de adultos fumantes, principalmente na categoria de ”fuma atualmente”, e uma predominância de pacientes casados entre os idosos. No dataset de Diabetes, a maioria das pessoas estudadas não possui a doença, e identificamos uma relação positiva entre os níveis de glicose no sangue e a presença de diabetes. A análise de correlação confirmou

essas associações entre as variáveis, e gráficos de dispersão e box plots mostraram uma distribuição mais elevada de glicose nos pacientes com diabetes. Essas descobertas são essenciais para o desenvolvimento de estratégias de prevenção, diagnóstico precoce e tratamento eficaz, e a visualização dos dados facilita a compreensão e comunicação dos resultados para profissionais de saúde.

REFERENCES

[1] BEHRMAN, Kennedy R. ”Fundamentos de Python para ciência de dados”. Porto Alegre: Grupo A, 2023. E-book. ISBN 9788582605974. Disponível em: https://app.minhabiblioteca.com.br/#/books/9788582605974/. Acesso em: 21 jun. 2023.

[2] FÁVERO, Luiz P. ”Análise de Dados”. Rio de Janeiro: Grupo GEN, 2015. E-book. ISBN 9788595153226. Disponível em: https://app.minhabiblioteca.com.br/#/books/9788595153226/. Acesso em: 01 jul. 2023.

[3] NETTO, Amilcar; MACIEL, Francisco. ”Python para Data Science e Machine Learning Descomplicado”. Rio de Janeiro: Editora Alta Books, 2021. E-book. ISBN 9786555203172. Disponível em: https://app.minhabiblioteca.com.br/#/books/9786555203172/. Acesso em: 28 jun. 2023.