Real Data-Driven Business Project

Análise de Dados de Saúde para Previsão de Doenças

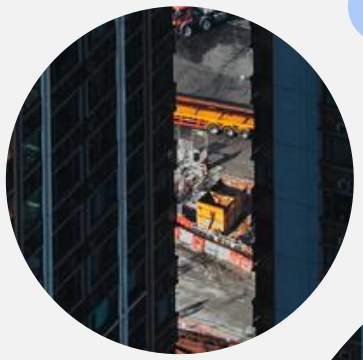
2401327 - Allison Melo dos Santos
2401077 - Bruno Rodrigues Martins
2402582 - Douglas Cardoso dos Santos
2402342 - Gislaine Gomes dos Santos
2402481 - Tatiana Araújo Domingos

07 de Outubro de 2024

MBA_DSA_29 - Teams 06

Agenda

- Teams 06;
- Trilha - Trabalho;
- Introdução;
- Arquitetura – Teórica / Técnica;
- Desenvolvimento e Análises;
- Conclusão;
- Considerações;
- Agradecimentos;



TEAMS 06

Allison Melo dos Santos

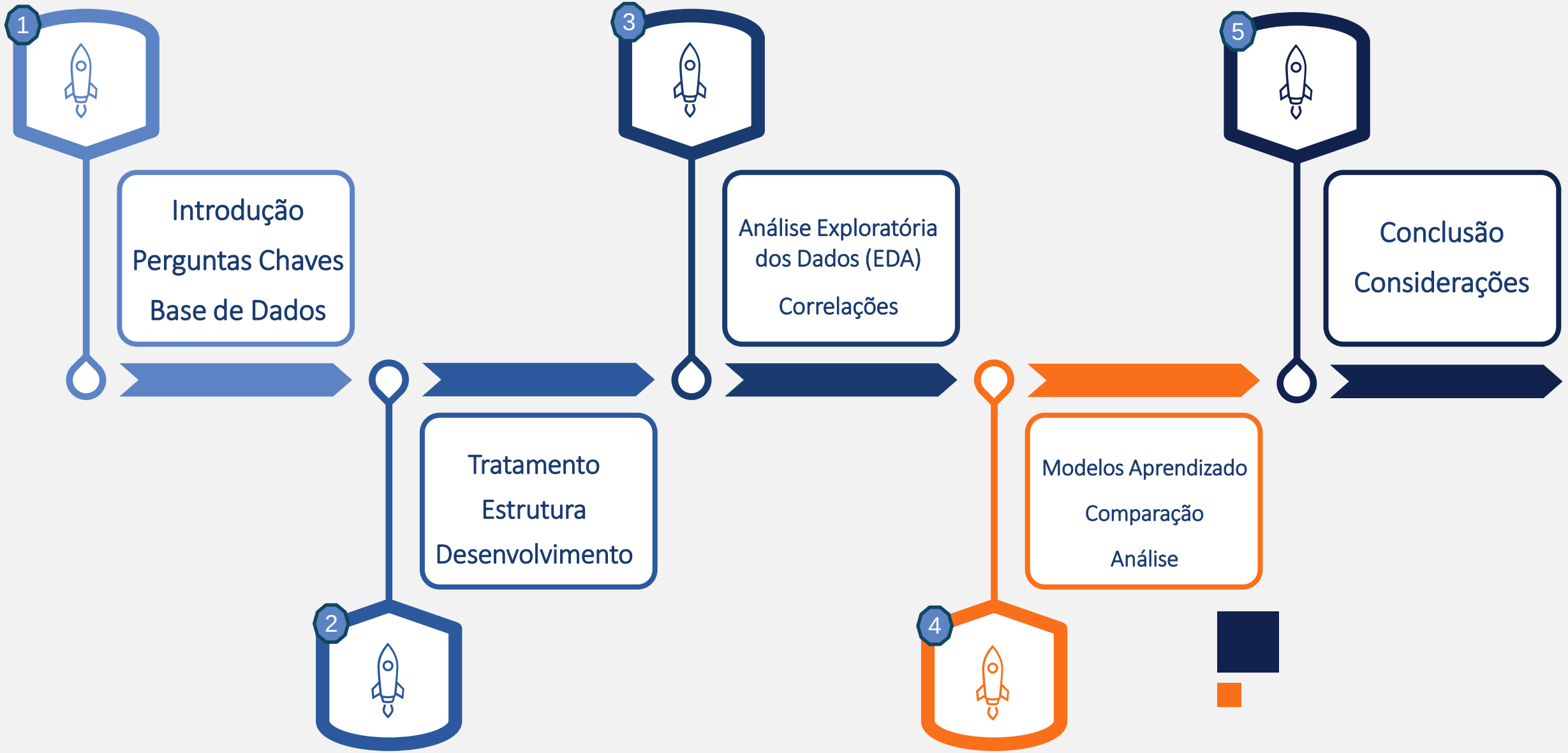
Bruno Rodrigues Martins

Douglas Cardoso dos Santos

Gislaine Gomes dos Santos

Tatiana Araujo Domingos

TRILHA - TRABALHO



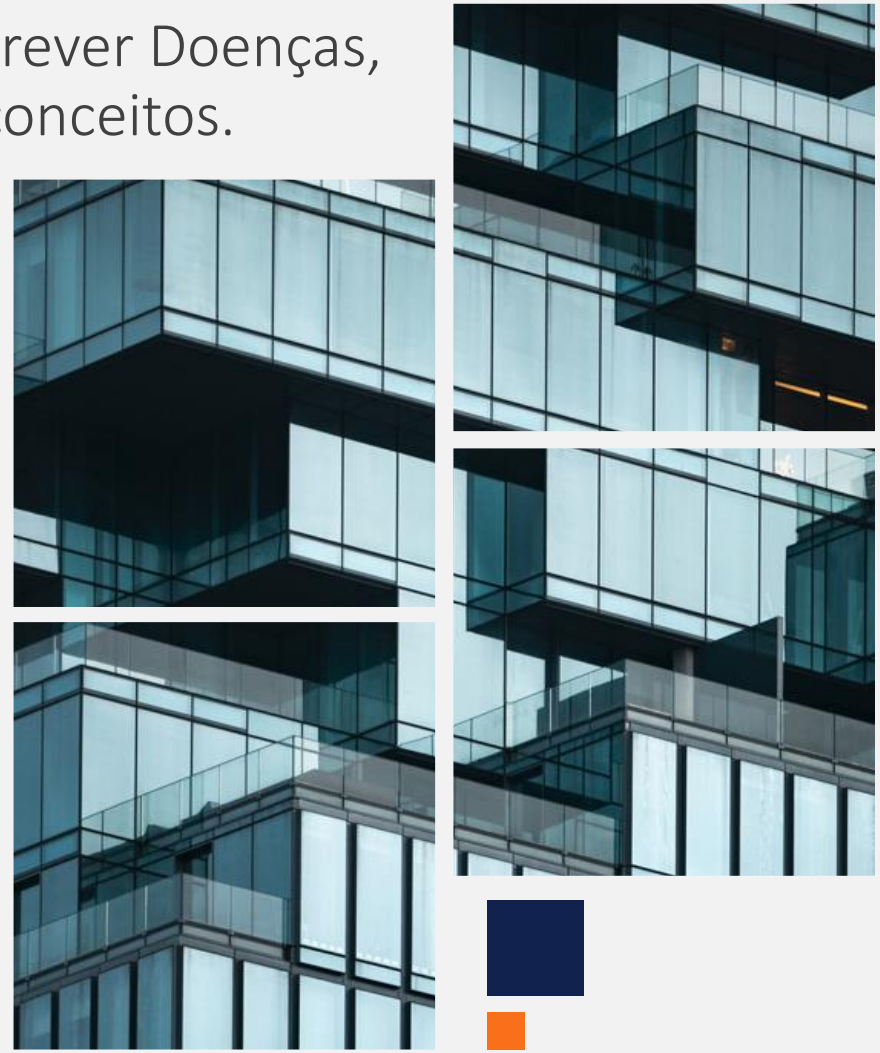
Introdução

Trabalho Desenvolvido para Análise de Dados e Prever Doenças, com aplicação de ferramentas, métricas e conceitos.

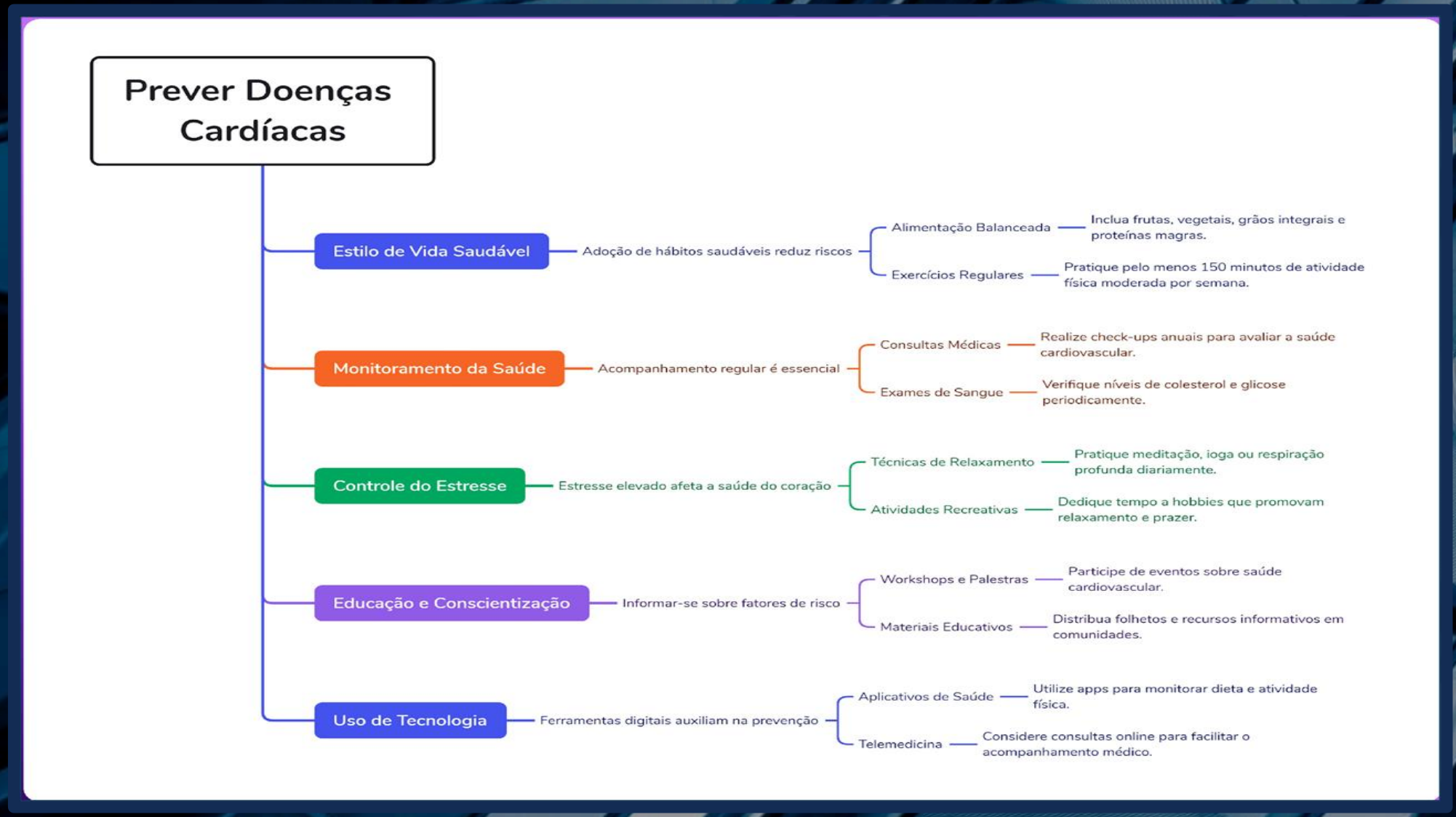
- Área de Estudo - Cardiologia

Objetivo: Identificar e prever o risco de doenças cardíacas utilizando dados históricos e características dos pacientes.

- Questões a Serem Respondidas:
  - Quais são os principais fatores de risco para doenças cardíacas?
  - Como podemos prever a probabilidade de um paciente desenvolver uma doença cardíaca com base em seus dados?



Introdução

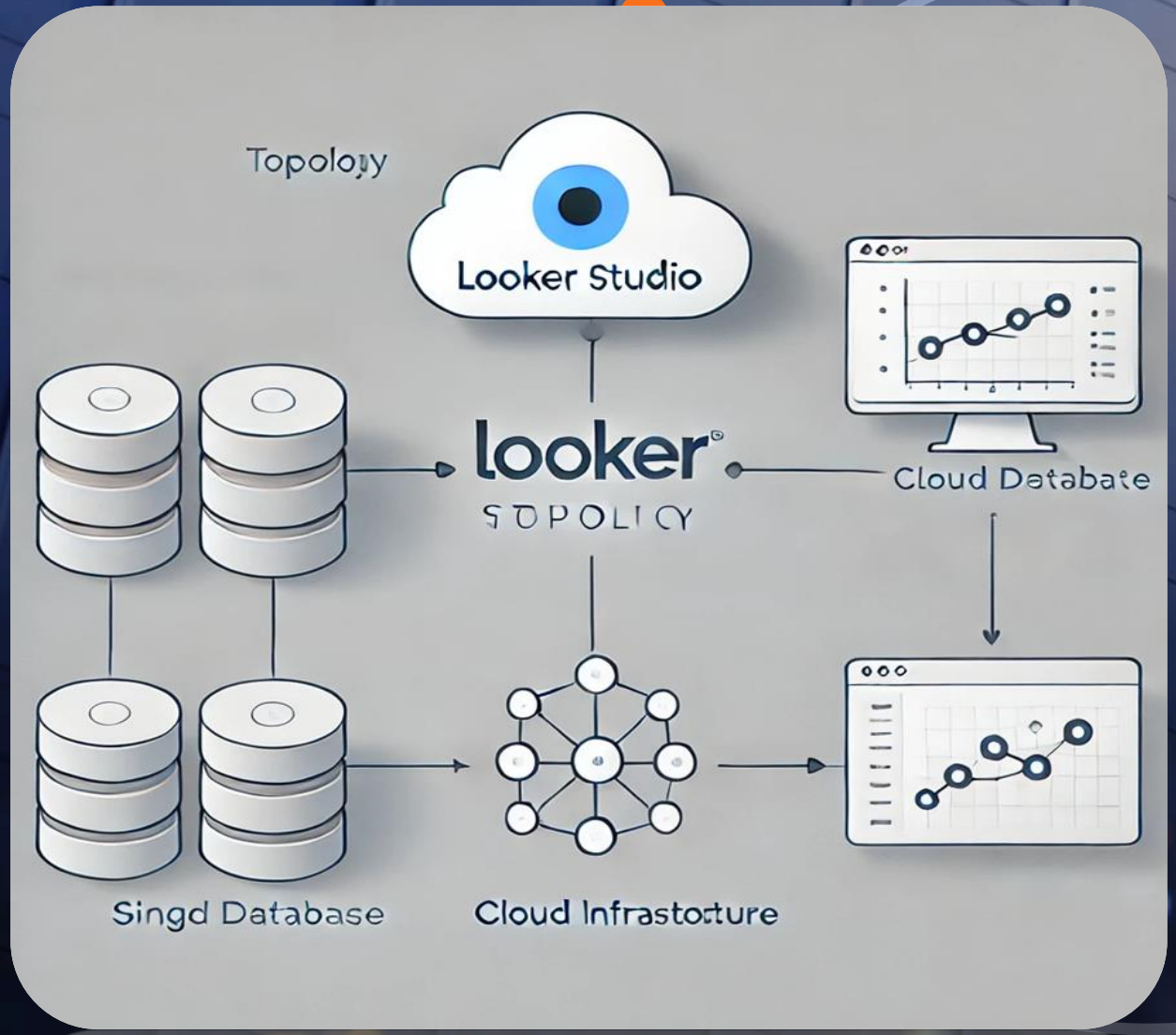


Mapa Mental – Como prever Doenças Cardíacas

Arquitetura - Teórica

A estrutura teórica para a análise de prevenção de doenças cardíacas pode ser organizada em diversas camadas, incorporando dados clínicos, processamento sofisticado de informações e produção de percepções úteis.

Arquitetura - Técnica



A definição da Infraestrutura, teve como tomada de decisão, baseado no prévio conhecimento de um membro que trabalha com Google Looker Studio.

Baseado nessa escolha, a aplicação do Ecossistema da Google Cloud possibilitando a implementação da solução e posteriormente recursos para integração com fonte de Dados como Hospitais, Clínicas de Exames, Aplicativos e Formulários cadastrados pelos pacientes.

Também utilizamos Notebook no Colab para desenvolvimento do trabalho.

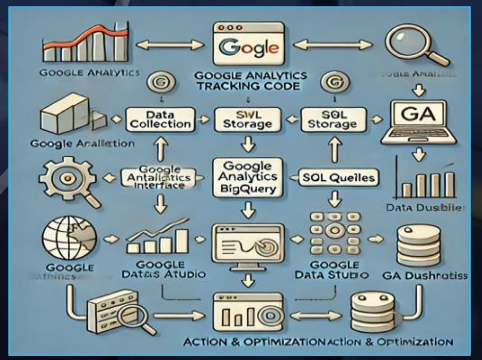
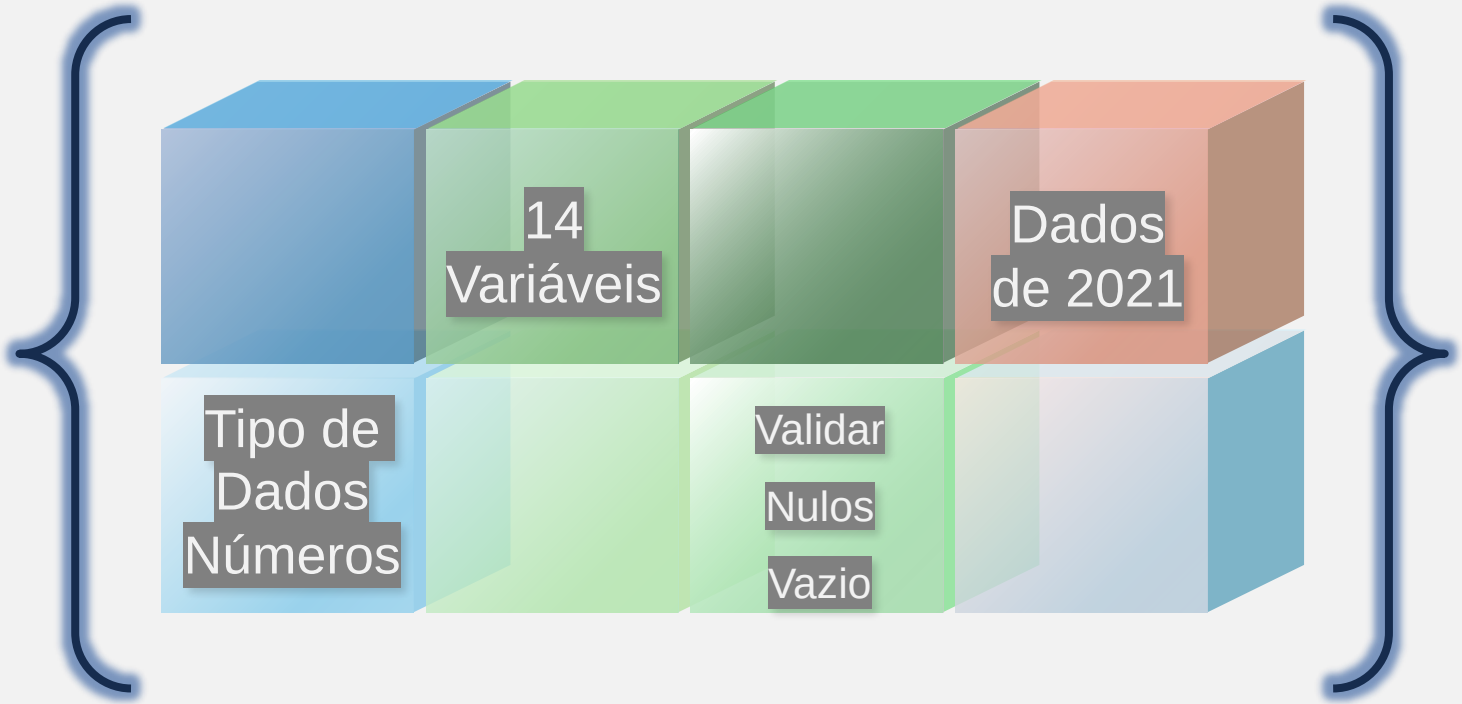


Figura 1 e 2: https://files.oaiusercontent.com/file-qv9pxEYFek87YXiuC11XEfLv?se=2024-10-06T18%3A36%3A37Z&sp=r&sv=2024-08-04&sr=b&rscc=max-age%3D604800%2C%20immutable%2C%20private&rscd=attachment%3B%20filename%3D083745cb-5aea-4156-8901-0e0661f7dc68.webp&sig=/ZuOP/j/GuNQQ9kgRiI2K/cRzKiKZU1GKS1k8GD5oNg%3D

BASE de DADOS

Descritivo

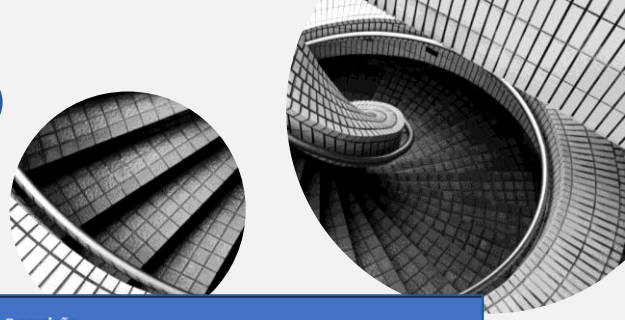
heart.csv



Dataset: https://raw.githubusercontent.com/Piggk2/RealData_DSA/refs/heads/main/heart.csv

Desenvolvimento e Análises

Dicionário



A **compreensão** e **interpretação** destas **variáveis** são cruciais na **previsão** e **diagnóstico** de doenças cardíacas, e podem orientar futuras **investigações** e **decisões** de **tratamento**.

Coluna (Variável)	Nome	Descrição
age	Age	Este é um fator de risco chave para doenças cardíacas. À medida que a idade aumenta, o risco de artérias danificadas e estreitadas, músculo cardíaco enfraquecido ou espessado e outros fatores de risco de doenças cardíacas também aumenta.
sex	Sex	Os homens geralmente correm maior risco de doenças cardíacas do que as mulheres. No entanto, após a menopausa, o risco da mulher aumenta quase igualando o do homem.
cp	Chest Pain Type (cp) "Tipo de dor no peito"	A dor no peito é um sintoma chave de doença cardíaca. Pode manifestar-se de diferentes formas: angina típica, angina atípica, dor não anginosa ou até mesmo ser assintomática. A dor no peito associada a doenças cardíacas é geralmente descrita como desconforto, peso, pressão, dor, queimação, plenitude, aperto ou sensação dolorosa.
trtbps	Resting Blood Pressure (trtbps) "Pressão arterial em repouso"	A pressão alta (hipertensão) pode endurecer e engrossar as artérias, levando ao acúmulo de placas (aterosclerose) que podem causar doença arterial coronariana. A pressão é medida em milímetros de mercúrio (mm Hg) e geralmente é registrada como dois algarismos. A pressão arterial normal em repouso em um adulto é de aproximadamente 120/80 mm Hg.
chol	Serum Cholesterol (chol) "Colesterol Ruim"	O colesterol é um tipo de molécula lipídica. Altos níveis de lipoproteína de baixa densidade (LDL) ou "colesterol ruim" podem aumentar o risco de doenças cardíacas, formando placas e estreitando as artérias.
fbs	Fasting Blood Sugar (fbs) "Níveis de Açúcar no Sangue"	Níveis elevados de açúcar no sangue em jejum (pré-diabetes ou diabetes) podem contribuir para o estreitamento das artérias e aumentar o risco de doenças cardíacas. Um nível de açúcar no sangue em jejum inferior a 100 mg/dL é considerado normal. 100-125 mg/dL é considerado pré-diabetes, e 126 mg/dL ou superior em dois testes separados significa que você tem diabetes.
restecg	Resting Electrocardiographic Results (restecg) "Resultados eletrocardiográficos em repouso"	O ECG registra a atividade elétrica do coração e pode mostrar ataques cardíacos anteriores ou problemas de ritmo cardíaco. Resultados anormais podem indicar problemas cardíacos, como hipertrofia ventricular esquerda ou arritmias cardíacas.
thalachh	Maximum Heart Rate Achieved (thalachh) "Frequência cardíaca máxima alcançada"	Durante exercícios ou testes de esforço, a frequência cardíaca máxima pode indicar a aptidão cardiovascular e a capacidade do coração de suportar o esforço.
exng	Exercise Induced Angina (exang) "Angina induzida por exercício"	Isso acontece quando o músculo cardíaco não recebe tanto sangue (e, portanto, oxigênio) quanto necessita para o nível de atividade física, causando dor ou desconforto no peito.
oldpeak	ST Depression Induced by Exercise Relative to Rest (oldpeak) "Depressão ST induzida por exercício em relação ao repouso"	Alterações no segmento ST em um ECG podem indicar doença cardíaca. A depressão do segmento ST pode indicar isquemia ou falta de fluxo sanguíneo suficiente para o músculo cardíaco.
slp	The Slope of The Peak Exercise ST Segment (slp) "A inclinação do segmento ST do pico do exercício"	A inclinação do segmento ST/frequência cardíaca (inclinação ST/FC) foi introduzida como um índice de demanda relativa de oxigênio do miocárdio durante o exercício. O formato do segmento ST pode revelar muito sobre a condição do coração
caa	Number of Major Vessels Colored by Flourosopy (caa) "Número de vasos principais coloridos pela Flourosopy"	Isso mede a presença de doenças nos principais vasos sanguíneos do coração. Um número maior geralmente indica doença mais grave.
thall	Thallium Stress Test (thall) "Teste de estresse com tálio"	Este é um método de imagem nuclear que mostra quão bem o sangue flui para o músculo cardíaco, tanto em repouso quanto durante a atividade. Pode revelar áreas do músculo cardíaco que não estão recebendo sangue suficiente, indicando doença arterial coronariana.
output	Output (Diagnosis of Heart Disease) " Condição - Diagnóstico de doenças cardíacas"	Esta é a variável de destino. Um valor de 0 indica menos de 50% de estreitamento do diâmetro - não uma doença cardíaca significativa, enquanto um valor de 1 indica mais de 50% de estreitamento do diâmetro - uma doença cardíaca significativa.

Desenvolvimento e Análises

Análise Exploratória de Dados

Na Análise exploratória de dados (EDA), **identificamos** a **distribuição** de **características** individuais e analisando as **conexões entre elas**. Esta **fase** da análise é **vital** para **compreender** a **estrutura e as particularidades** dos nossos dados, possibilitando-nos fazer escolhas conscientes nos **próximos passos da análise.**

- ✓ Análise por Gêneros
- ✓ Análise por faixa etária;
- ✓ Análise por Atividades Físicas;

Dash: https://lookerstudio.google.com/u/2/reporting/eb3d5d09-7ce9-42c2-9c76-18792425288c/page/OllDE

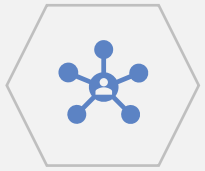
Desenvolvimento e Análises

Modelo Correlação

A **correlação** é um conceito **fundamental em estatística e análise de dados**, amplamente utilizado em diversas áreas para identificar e quantificar a relação entre variáveis, oferecendo insights valiosos para a tomada de decisão e a construção de modelos preditivos.

Grau de Correlação

- Pode ser medido pelos chamados coeficientes de correlação, sendo assim quando se tem uma correlação entre dois eventos , isso pode não demonstrar uma conexão casual entre eles.



Foco

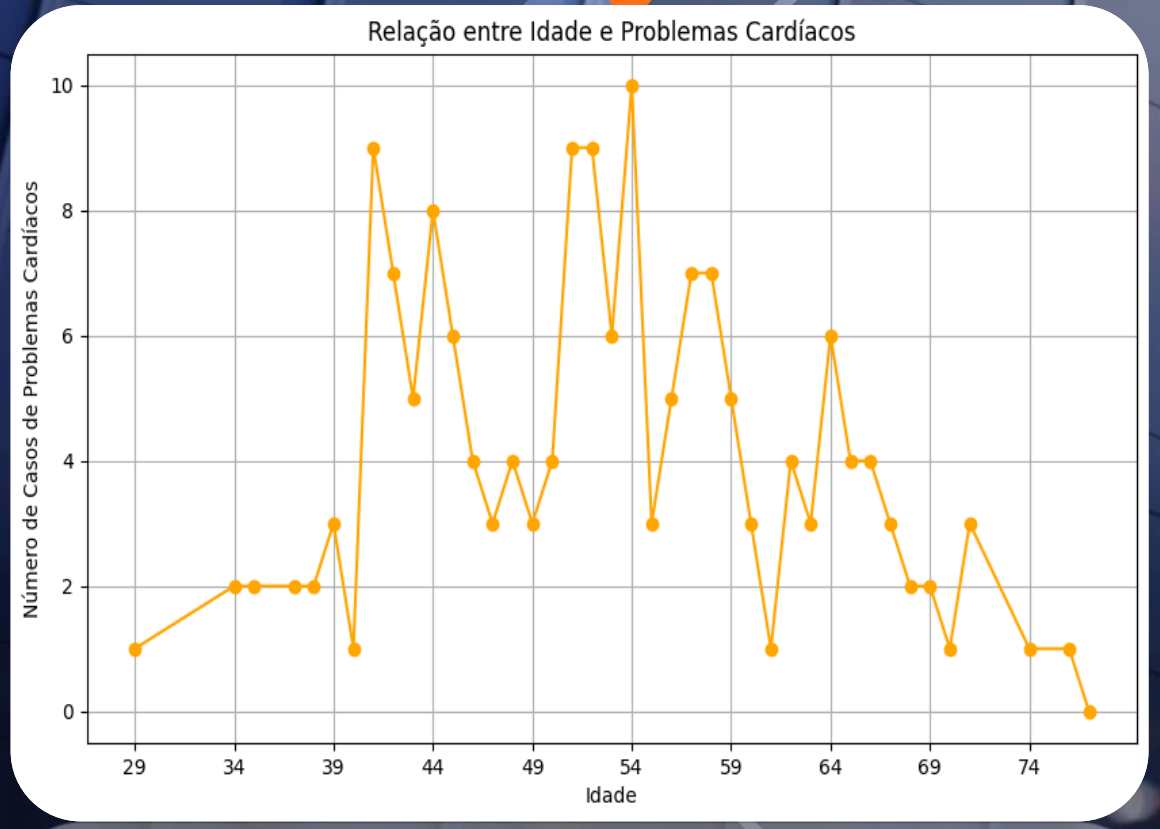
- Ferramenta crucial na análise de dados, permitindo entender melhor as relações entre variáveis e guiar decisões baseadas em dados. Para que a tomada de decisão seja de forma mais assertiva e com um maior ganho e também amplo conhecimento sobre um problema ou hipótese.

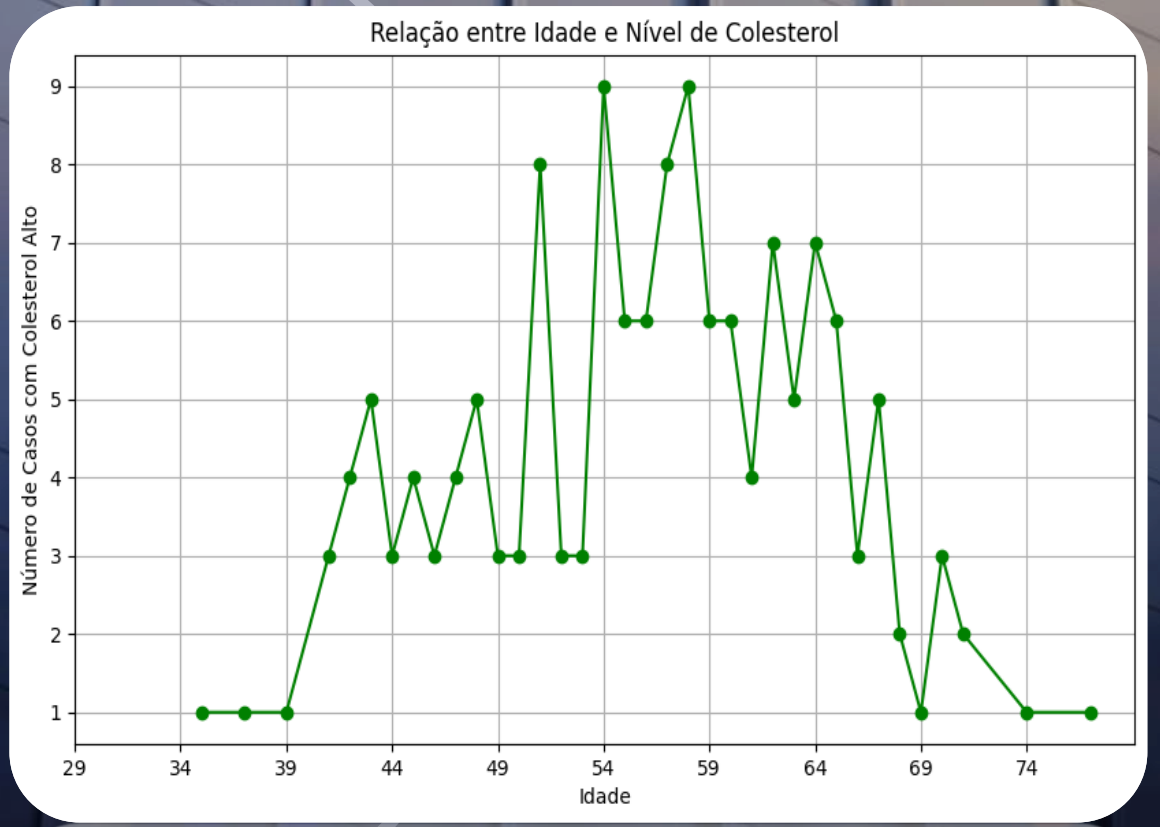
Identificação de Padrões e Relações

- Identificar se existe uma relação entre duas variáveis, permitindo descobrir padrões escondidos nos dados.
- Por exemplo, ao analisar dados de saúde, pode-se verificar se há uma correlação entre a idade e a incidência de uma doença, o que pode ser relevante para a prevenção.

Desenvolvimento e Análises

Tipos de Correlação



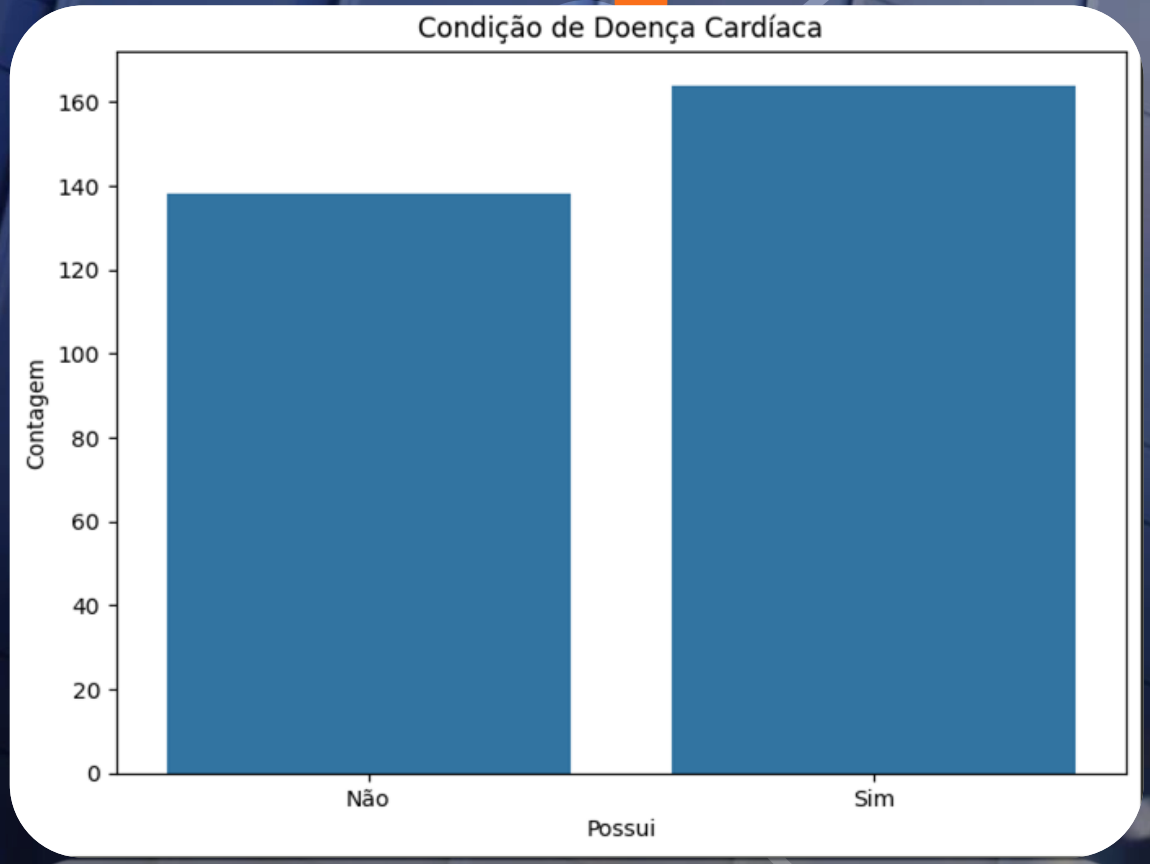


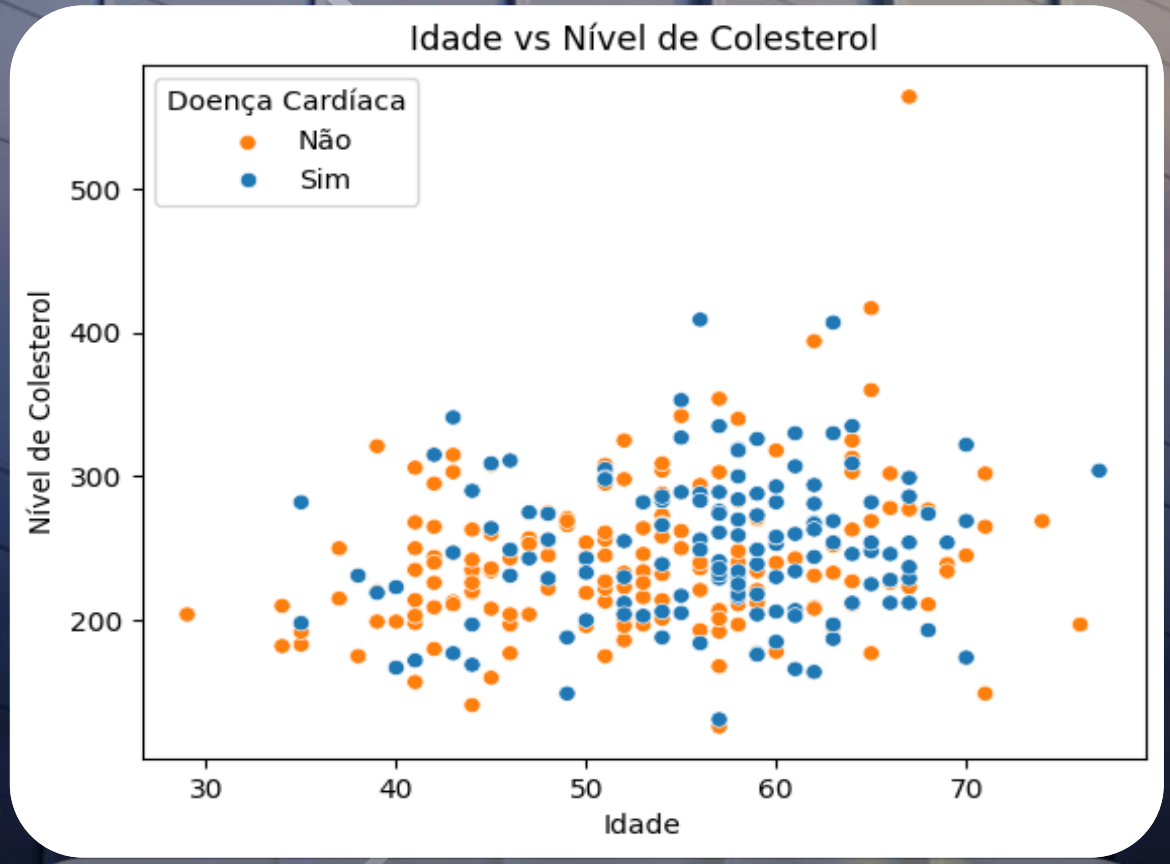
De acordo com analise realizada, o maior volume de casos de problemas cardíacos ocorrem na faixa etária de 50 a 60

Observou-se que a maior nível de colesterol foi encontrado entre pacientes de 50 a 60 anos.

Desenvolvimento e Análises

Tipos de Correlação





Observa-se que a quantidade de pessoas com doenças cardíacas e maior que para não possuem.

Pode se dizer que pessoas entre 40 a 55 anos apresentam um nível elevado de colesterol e consequentemente houveram problemas cardíacos.

Desenvolvimento e Análises

Modelo Machine Learning

Nosso objetivo era construir modelos de previsão de doenças cardíacas e comparar o desempenho de 3 algoritmos de aprendizado de máquina. O desempenho dos modelos foi avaliado com base na precisão de treinamento e teste.

Random Forest

- Cada árvore na floresta faz uma previsão, e o modelo escolhe a previsão;
- Poderoso por sua capacidade de gerar previsões precisas enquanto minimiza erros e generaliza melhor para novos dados.

Logistic Regression

- Apesar de ter "regressão" no nome, a regressão logística é um modelo de classificação.
- Utilizado para problemas de classificação binária

Decision Tree

- Funciona construindo um modelo em formato de árvore, onde cada nó representa uma decisão baseada em um atributo dos dados, e cada ramo representa o resultado dessa decisão

Modos de Avaliação

- Matriz de Confusão
- Relatório com Métricas.

Desenvolvimento e Análises

Logistic Regression

Classification Report for LogisticRegression:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	29
1	0.90	0.84	0.87	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

Boa Performance: Os resultados sugerem que esse modelo é eficaz na tarefa de classificação

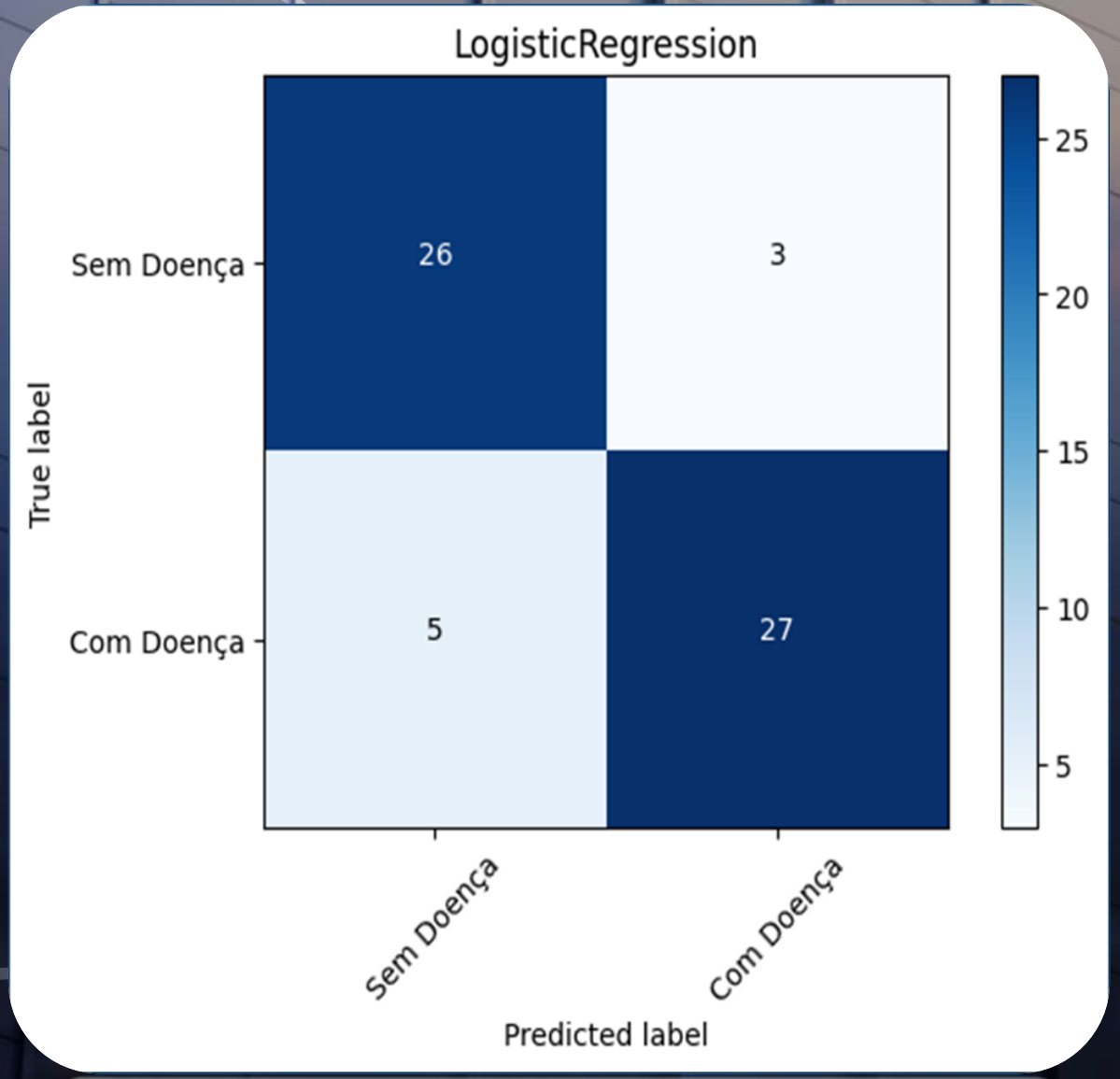
Métricas

- **Recall / Precision** - 87%
- **F1-score –** Mostra Equilíbrio das Métricas

Matriz de Confusão

- Valores estão balanceados

Ponto de preocupação : Por ser uma base não muito grande, temos um Support Alto, mas em datasets pequenos, é importante considerar que variações nas previsões podem impactar significativamente essas métricas.



Desenvolvimento e Análises

Decision Tree

Classification Report for DecisionTreeClassifier:

	precision	recall	f1-score	support
0	0.83	0.83	0.83	29
1	0.84	0.84	0.84	32
accuracy			0.84	61
macro avg	0.84	0.84	0.84	61
weighted avg	0.84	0.84	0.84	61

Boa Performance: Os resultados sugerem que esse modelo é eficaz na tarefa de classificação

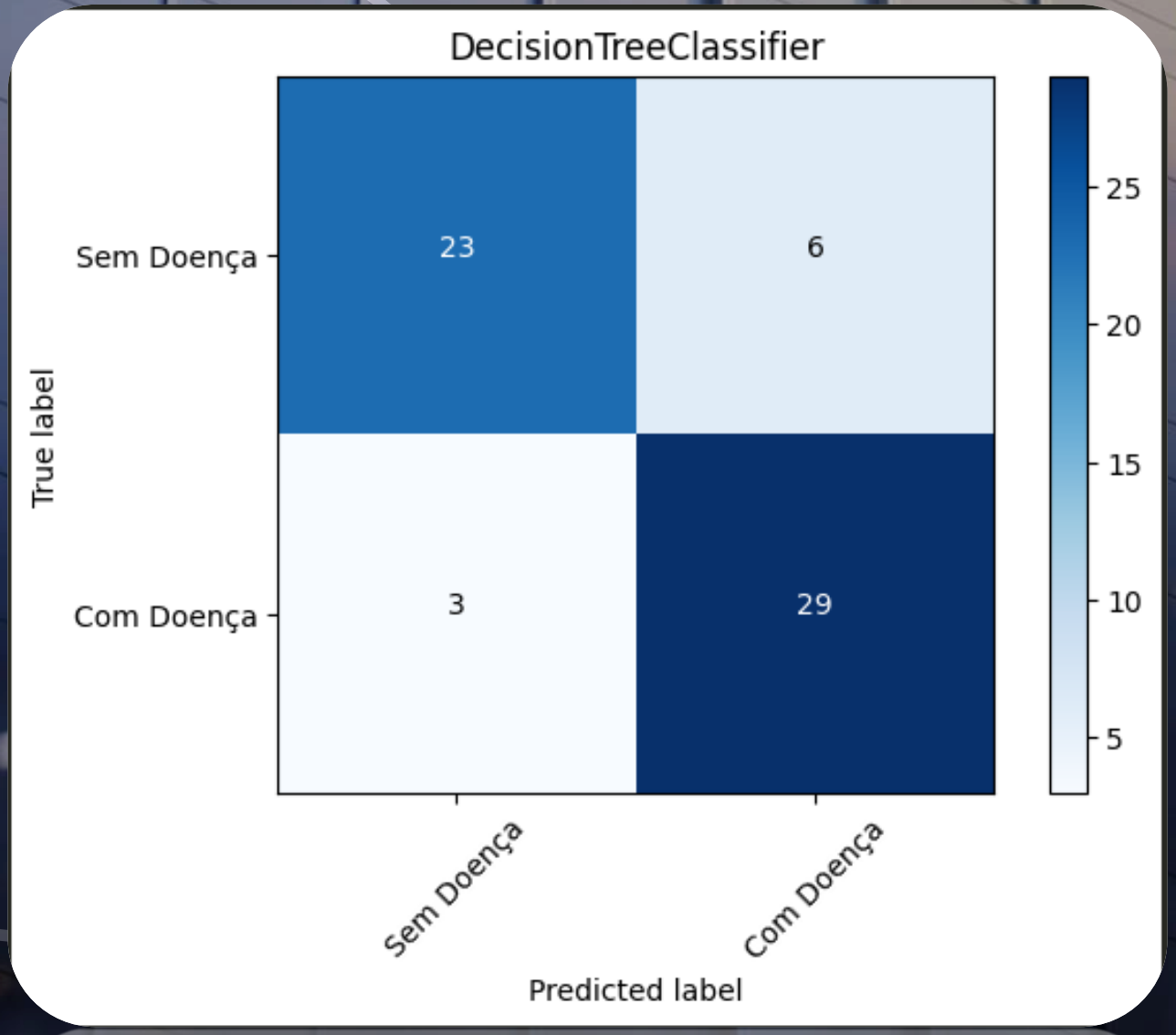
Métricas

- **Recall / Precision** - 84%
- **F1-score –** Mostra Equilíbrio das Métricas

Matriz de Confusão

- Valores estão balanceados

Ponto de preocupação : Por ser uma base não muito grande, temos um Support Alto, mas em datasets pequenos, é importante considerar que variações nas previsões podem impactar significativamente essas métricas.



Desenvolvimento e Análises

Random Forest

Classification Report for RandomForestClassifier:

	precision	recall	f1-score	support
0	0.87	0.90	0.88	29
1	0.90	0.88	0.89	32
accuracy			0.89	61
macro avg	0.88	0.89	0.89	61
weighted avg	0.89	0.89	0.89	61

Boa Performance: Os resultados sugerem que esse modelo é eficaz na tarefa de classificação

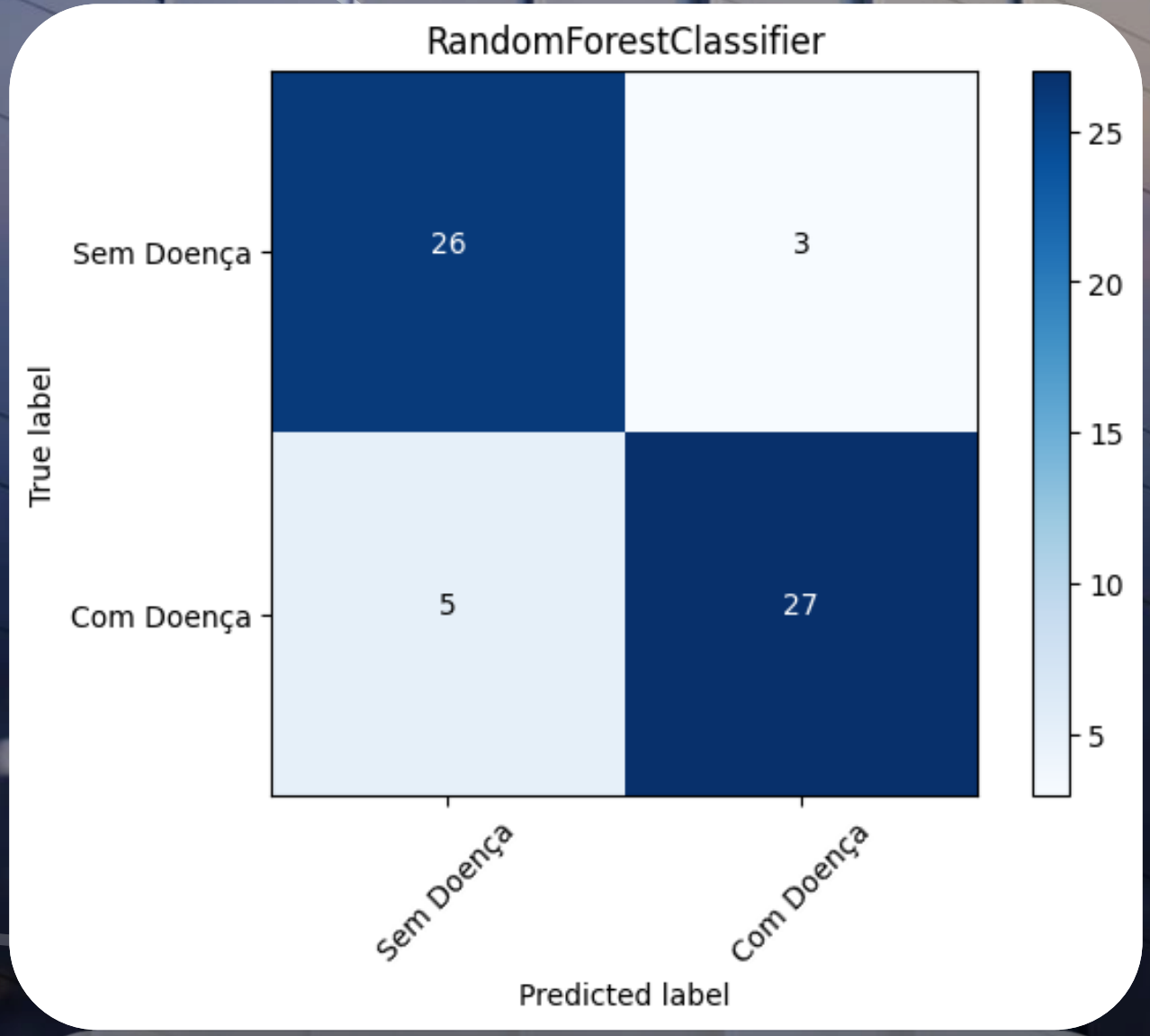
Métricas

- **Recall / Precision** - 89%
- **F1-score –** Mostra Equilíbrio das Métricas

Matriz de Confusão

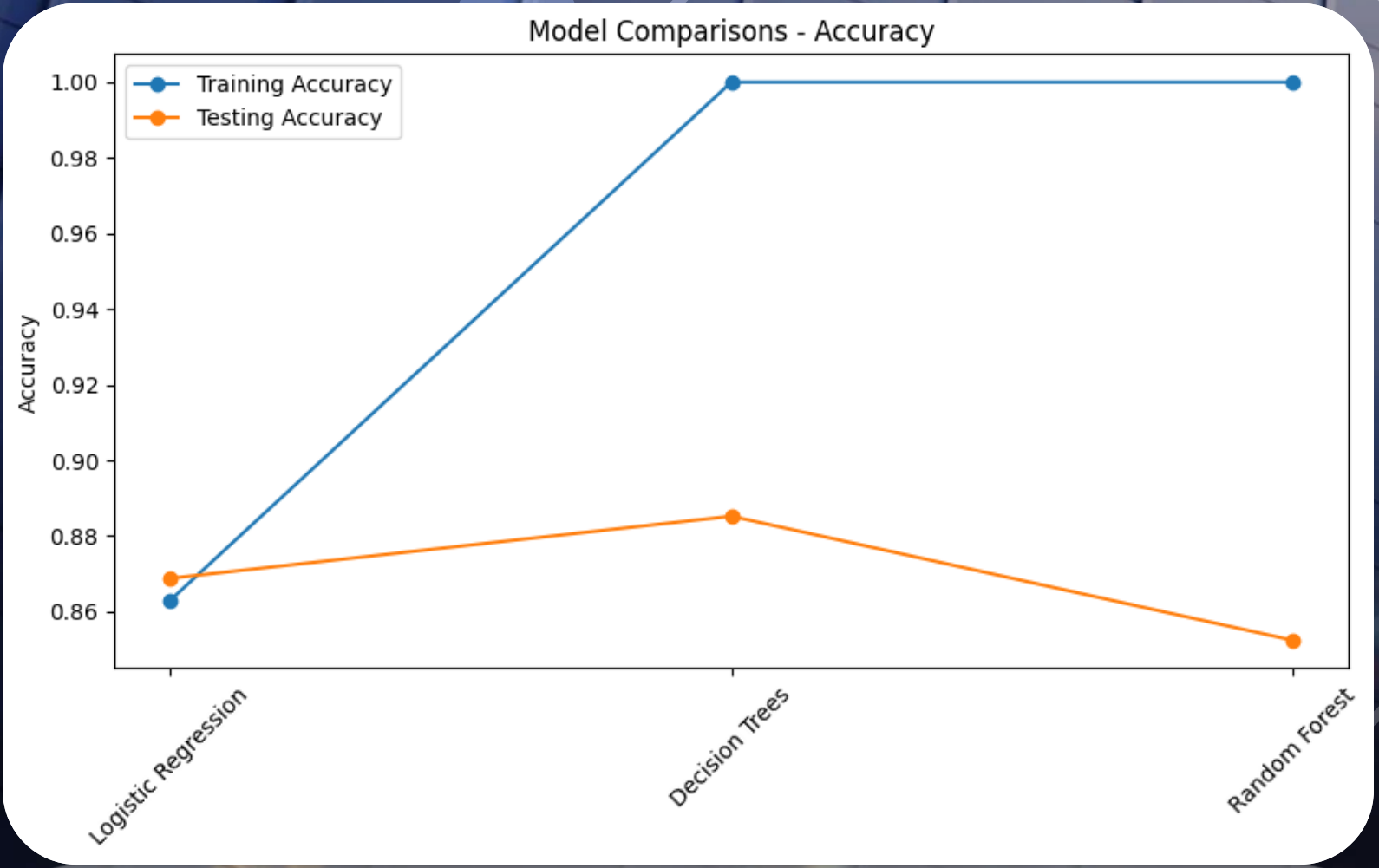
- Valores estão balanceados

Ponto de preocupação : Por ser uma base não muito grande, temos um Support Alto, mas em datasets pequenos, é importante considerar que variações nas previsões podem impactar significativamente essas métricas.



Desenvolvimento e Análises

Comparação



A acurácia é uma medida frequentemente utilizada para avaliar o desempenho de modelos de classificação.

No comparativo ao lado vemos que 2 modelos na base de treino é 100%, que indica que o modelos são eficazes com o conjunto de dados, após o aprendizado.
Mas na base de teste com menor conjunto de dados a eficácia e menor.

Vemos que o modelo de Logistic Regression possui % de acucária mais próxima entre teste /treino.

Conclusão

Análise Exploratória de Dados (EDA):
Na fase EDA, tivemos uma noção da distribuição dos dados e examinamos as propriedades estatísticas básicas dos dados. Isso incluiu a compreensão dos tipos de variáveis, a verificação de valores ausentes e a visualização da distribuição de vários recursos e da variável de destino. Esta etapa inicial é crucial em qualquer projeto de ciência de dados e nos ajudou a identificar tendências, anomalias, padrões e relacionamentos nos dados.

Análise de Correlação:
A seguir, realizamos uma análise de correlação para compreender as relações entre diferentes características numéricas.

Previsão de aprendizado de máquina:
Aplicamos 3 modelos diferentes de aprendizado de máquina para previsão de doenças cardíacas e avaliamos seu desempenho em termos de precisão. E comparação entre eles.

A aplicabilidade da Análise de dados como nesses estudo podemos ter cenários: Gestão de Planos de Saúde / Hospitais para ter mais capacidades de atendimento / Industrias farmacêuticas quanto a desenvolvimento de remédios / Academias para possibilitar diminuir os Riscos, etc.

Conclusão

Resumo dos Resultados

Os principais fatores de risco para doenças cardíacas identificados foram idade, níveis de colesterol e atividades físicas.
A faixa etária de maior risco é entre 50 a 60 anos, com uma alta ligação entre níveis altos de colesterol e incidência de doenças cardíacas.

Desempenho dos modelos preditivos

Entre os modelos mostrados, o Random Forest apresentou a melhor precisão para a previsão de doenças cardíacas, seguido pelo Decision Tree e Logistic Regression.
O Modelo Random Forest, indica sua capacidade de distinguir entre pacientes com alto e baixo risco de desenvolver problemas cardíacos.

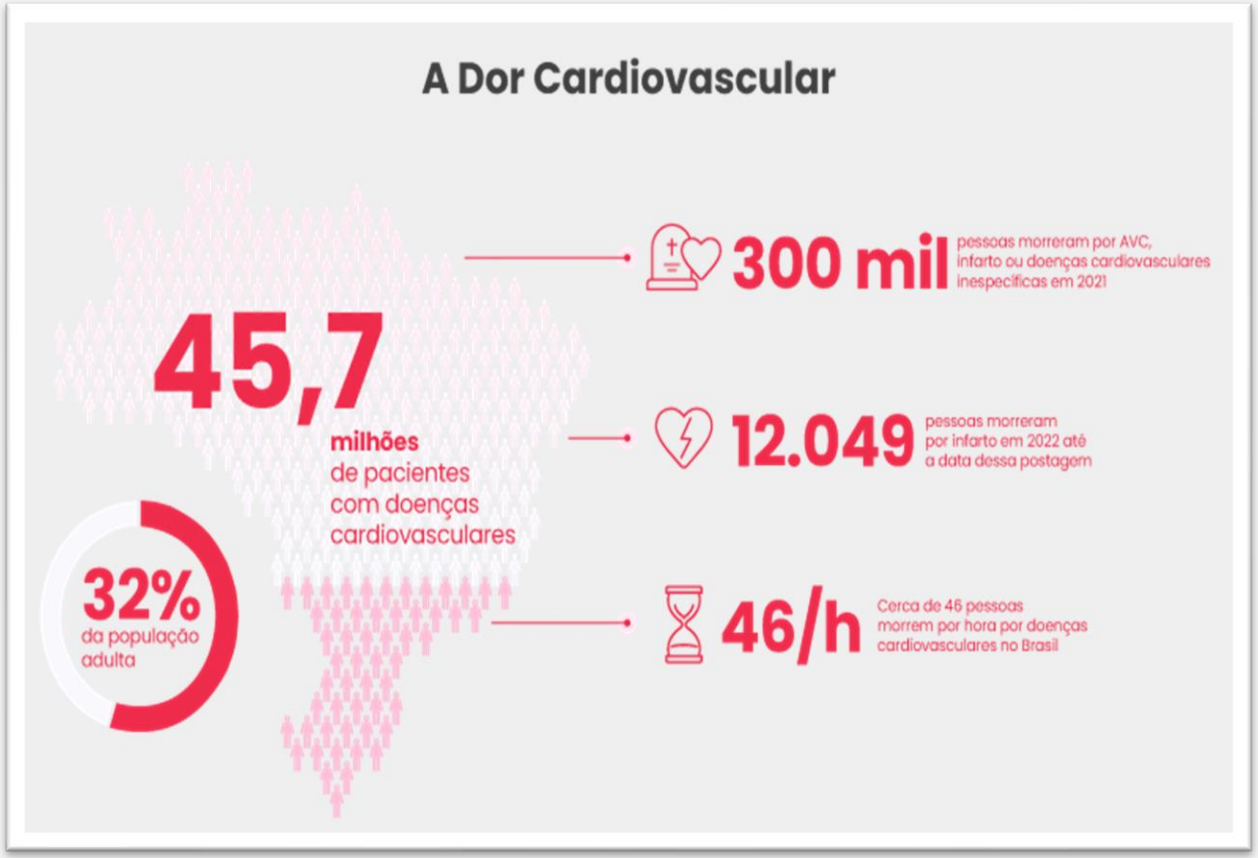
Potencial impacto das previsões

A identificação precoce de fatores de risco, como níveis elevados de colesterol, permite preventivas personalizadas, como controle alimentar e exercícios físicos.
A implementação programas preditivos em hospitais pode ajudar a priorizar atendimento e direcionar recursos para quem mais precisa.

Pontos a considerar

Este estudo foi realizado com uma base de dados limitada, e de apenas uma fonte. Realizar a ampliação da base pode aumentar a robustez e gerar modelos preditivos melhores.
A falta de variáveis como níveis de estresse, dieta e estilo de vida dos pacientes estudados pode ter limitado a identificação de todos os fatores de risco.

Considerações



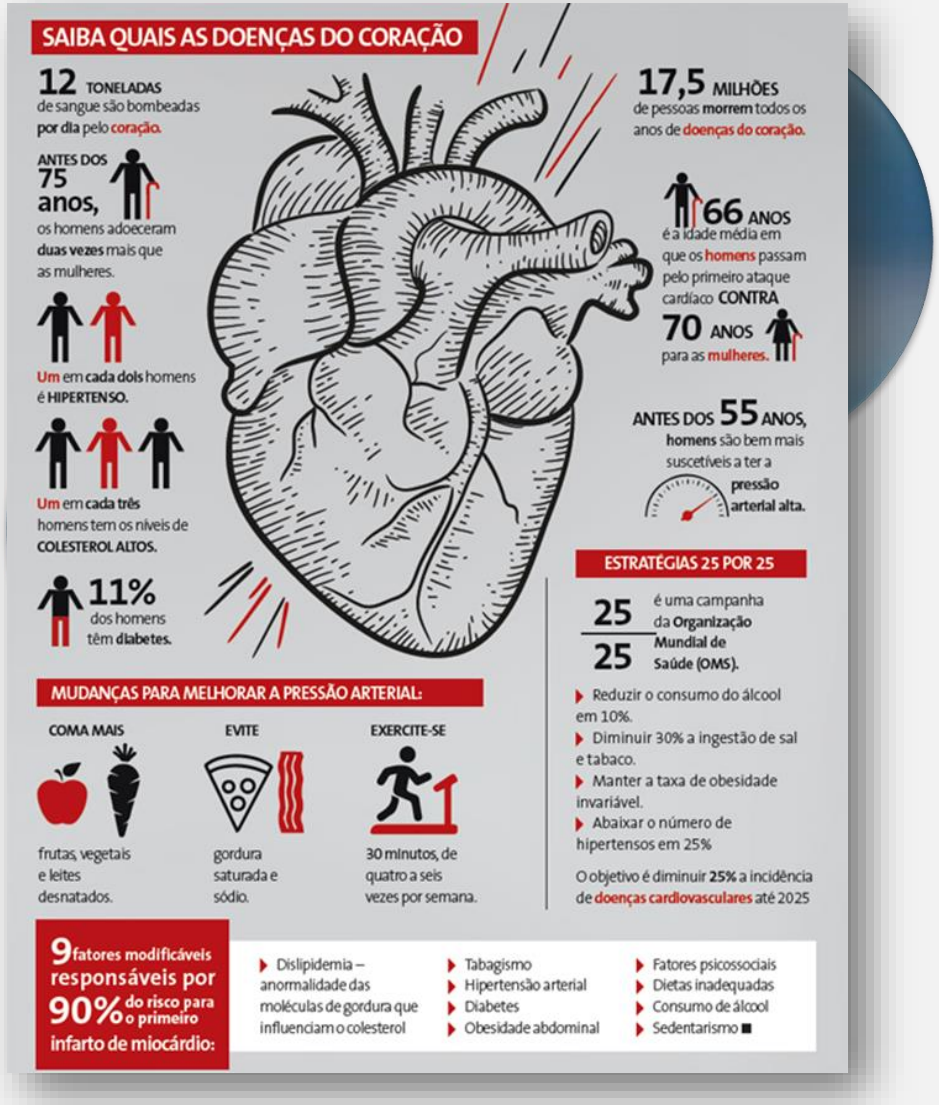


Figura 1: https://neomed.com.br/um-panorama-das-doencas-cardiovasculares-no-brasil/

Figura 2: https://www.pulsocardiologia.com.br/site/new.php?corpo=conteudo.php&tabela=tabram01&pg=1&cod=117

OBRIGADO!!!

Alisson, Bruno, Douglas, Gislaine, Tatiana