

HEART ATTACK

Uma análise do conjunto de dados

ENTENDIMENTO DO PROBLEMA

O dataset analisado é um banco de dados sintético e foi retirado da plataforma Kaggle.

Ele contém fatores de risco de ataque cardíaco relacionados à indivíduos da África do Sul. Inclui detalhes demográficos, histórico médico, hábitos de estilo de vida e medidas clínicas para avaliar resultados de ataque cardíaco.

- O conjunto de dados é projetado para modelagem preditiva, análise estatística e aplicações de aprendizado de máquina em pesquisa de saúde.

OBJETIVO DA ANÁLISE

Encontrar a resposta para a seguinte pergunta:

- Quais os principais fatores associados à incidencia de ataque cardíaco?

A resposta para essa pergunta pode direcionar estratégias de prevenção precoce.

CONHECENDO AS VARIÁVEIS

Nome	Explicação
Patient_ID	Identificador único do paciente.
Age	Idade do paciente em anos.
Gender	Gênero do paciente: Masculino ou Feminino.
Cholesterol_Level	Nível de colesterol total no sangue, medido em miligramas por decilitro (mg/dL).
Blood_Pressure_Systolic	Pressão arterial sistólica do paciente, medida em milímetros de mercúrio (mmHg).
Blood_Pressure_Diastolic	Pressão arterial diastólica do paciente, medida em milímetros de mercúrio (mmHg).
Smoking_Status	Status de tabagismo do paciente: Fumante (Sim) ou Não Fumante (Não).
Alcohol_Intake	Nível de consumo de álcool: Baixo, Moderado ou Alto.
Physical_Activity	Nível de atividade física do paciente: Sedentário, Ativo ou Altamente Ativo.
Obesity_Index	Índice de Massa Corporal (IMC), que é uma medida de obesidade calculada a partir do peso e altura.
Diabetes_Status	Status de diabetes do paciente: Tem diabetes (Sim) ou Não tem diabetes (Não).

CONHECENDO AS VARIÁVEIS

Family_History_Heart_Disease	Histórico familiar de doenças cardíacas: Sim (se houver histórico) ou Não (se não houver).
Diet_Quality	Qualidade da dieta do paciente: Ruim, Média ou Boa.
Stress_Level	Nível de estresse do paciente: Baixo, Médio ou Alto.
Heart_Attack_History	Histórico de infarto do miocárdio: Já teve infarto (Sim) ou Nunca teve (Não).
Medication_Usage	Uso de medicação pelo paciente: Sim (usa medicação) ou Não (não usa medicação).
Triglycerides_Level	Nível de triglicerídeos no sangue, medido em miligramas por decilitro (mg/dL).
LDL_Level	Nível de LDL (lipoproteína de baixa densidade), conhecido como "colesterol ruim" (mg/dL).
HDL_Level	Nível de HDL (lipoproteína de alta densidade), conhecido como "colesterol bom" (mg/dL).
Heart_Attack_Outcome	Resultado de um ataque cardíaco: 0 (Não) ou 1 (Sim), indicando se o paciente sofreu um infarto.

ETAPAS DA ANÁLISE

- Análise Descritiva, Univariada e Bivariada;
- Tratamento de dados;
- Análise da Correlação entre as Variáveis;
- Obtenção do Information Value;
- Treinamento de modelos;
- Avaliação dos Modelos através das Métricas: Acurácia, AUC, GINI e KS.
- Conclusões.

ANÁLISE DESCRIPTIVA

- Ausência de dados faltantes;
- Idade média dos pacientes é 56,9 anos, sendo que 25% do total tem idade igual a 73 anos ou mais e 25% tem 41 anos ou menos, indicando que a nossa amostra tem predominantemente adultos e idosos.
- As duas pressões analisadas (sistólica e diastólica) estão com a média acima do valor ideal, sugerindo prevalência de hipertensão;
- o Indice de Massa Corporal médio também sugere predominância de sobre peso na amostra.

ANÁLISE DESCRIPTIVA

- Média de Triglicerídeos de 174,6 mg/dL (ideal: <150 mg/dL), com 25% dos pacientes acima de 237 mg/dL, indicando um fator de risco significativo;
- Nível médio de colesterol acima do recomendado (<190 mg/dL), com 75% da amostra com um colesterol de 187 mg/dL ou mais.
- 58,7% da amostra analisada já sofreu ataque cardíaco.

Resumindo, os dados indicam um perfil de pacientes com hipertensão, colesterol elevado, obesidade e triglicerídeos altos.

ANÁLISE UNIVARIADA

- A distribuição de gênero está bem equilibrada;
- Temos uma quantidade muito maior de não fumantes;
- A maioria consome alcool de forma baixa ou moderada;
- A proporção de sedentários representa metade das amostras;
- O número de não diabéticos é praticamente 4x maior que o de diabéticos;
- A maioria não tem histórico de doenças cardíacas na família;
- A maioria tem uma dieta boa ou mediana;
- Metade dos pacientes relata ter estresse mediano;
- A proporção de pacientes que usam ou não medicamentos é a mesma;
- Como já mencionado, a maior porção já sofreu ataque cardíaco.

ANÁLISE BIVARIADA

- Fumantes tem maiores chances de ter ataque cardiaco em relação a não fumantes;
- O número de pacientes com diabetes que sofreram um infarto é consideravelmente maior do que aqueles que não sofreram e essa mesma relação se aplica aos que tem histórico familiar de ataque cardiaco;
- A idade das pessoas que sofreram ataque cardiaco é maior do que daquelas que não sofreram.

TRATAMENTO DE DADOS

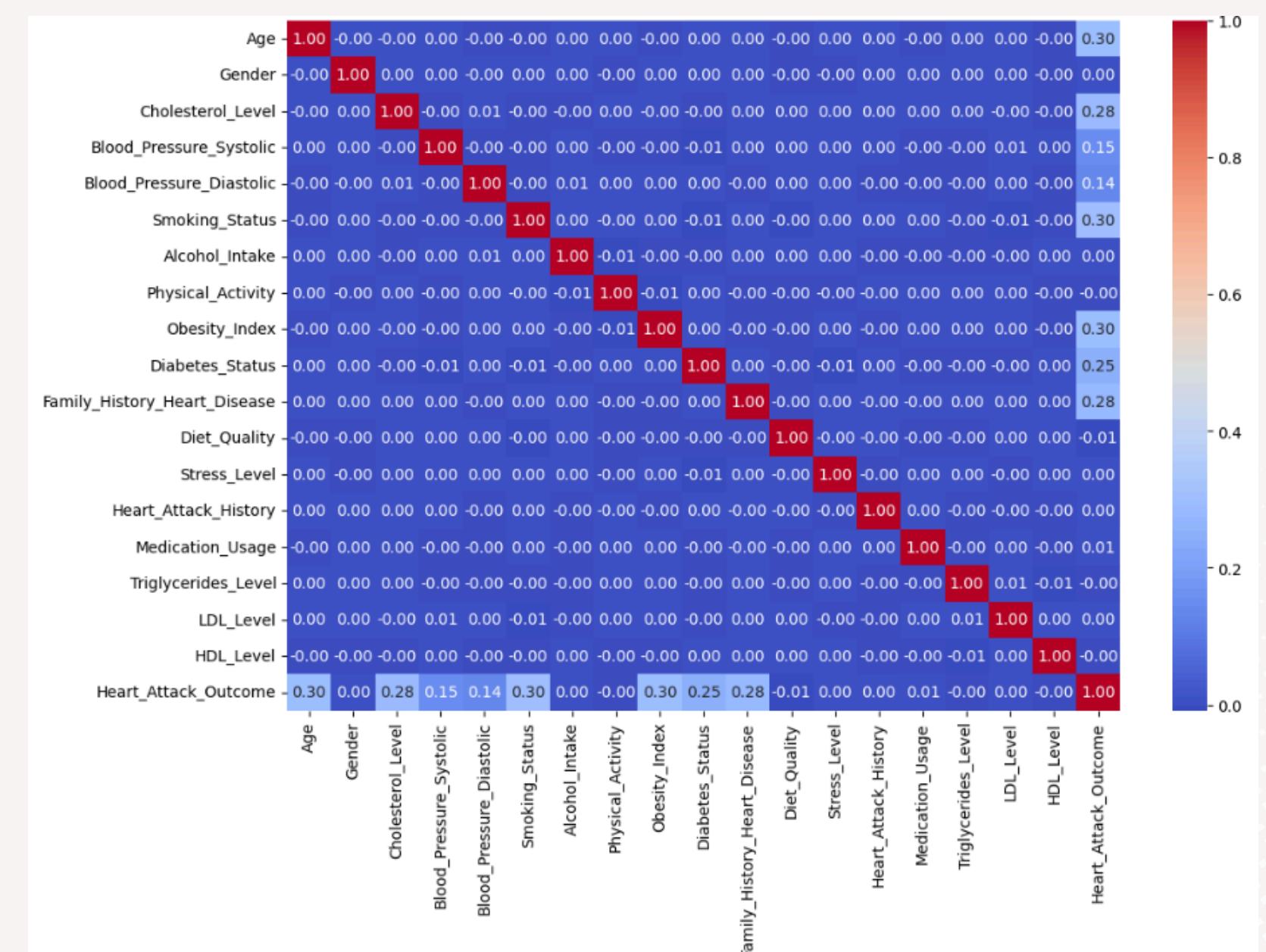
Como em nossos dados não haviam valores nulos e nem mesmo outliers, o único tratamento realizado foi quanto aos tipos de dados e a exclusão do número de identificação do cliente, uma vez que não possui poder preditivo.

Tinhamos 10 variáveis categóricas, como nenhuma delas tinha mais que 3 categorias, substituímos 0 para valores como ‘não’, ‘homem’, ‘baixo’, ‘sedentario’, 1 para ‘sim’, ‘mulher’, ‘medio’ e 2 para ‘alto’ e ‘bom’.

Assim, nosso dataframe ficou apenas com dados do tipo numérico.

ANÁLISE DE CORRELAÇÃO ENTRE AS VARIÁVEIS

Para analisarmos questões como correlação e multicolineariedade utilizamos a matriz de correlação e o heatmap abaixo:



INFORMATION VALUE (IV)

O IV é uma técnica que nos permite verificar o valor de informação de cada uma das nossas variáveis, quantificando a importância dessas variáveis na previsão de um resultado.

Geralmente considera-se um IV superior a 2%, no nosso caso, obtemos as seguintes variáveis como possíveis preditoras:

- Age;
- Cholesterol_Level;
- Blood_Pressure_Systolic;
- Blood_Pressure_Diastolic;
- Smoking_Status;
- Obesity_Index; Diabetes_Status;
- Family_History_Heart_Disease.

TREINAMENTO DOS MODELOS

Separamos os nossos dados em treino e teste, sendo 80% para treino, o que nos resultou em um tamanho final de 80 mil amostras e 20% para teste, ou 20 mil amostras.

TREINAMENTO DOS MODELOS

Como o nosso objetivo é prever zeros e uns, ou seja, uma classificação, incidência ou não incidência, utilizamos uma regressão logística, permitida pelo balançoamento que nossos dados oferecem.

Criamos 3 modelos de regressão logística:

- O primeiro com todas as variáveis;
- O segundo, com todas as variáveis contidas no primeiro modelo que ofereciam um p-value menor que 0,05;
- E o terceiro modelo, utilizando as 3 variáveis utilizadas como primeiros nós de decisão em uma árvore de classificação gerada.

TREINAMENTO DE MODELOS

A seguir, a equação de regressão utilizada em cada modelo. Lembrando que os modelos tem como objetivo realizar a previsão de sim e não para ataque cardíaco:

Todas as 18 variáveis (Modelo 1):

- *Heart_Attack_Outcome ~ Age + Gender + Cholesterol_Level + Blood_Pressure_Systolic + Blood_Pressure_Diastolic + Smoking_Status + Alcohol_Intake + Physical_Activity + Obesity_Index + Diabetes_Status + Family_History_Heart_Disease + Diet_Quality + Stress_Level + Heart_Attack_History + Medication_Usage + Triglycerides_Level + LDL_Level + HDL_Level*

TREINAMENTO DE MODELOS

As 8 variáveis com p-value menor que 0,05 (Modelo 2):

- $\text{Heart_Attack_Outcome} \sim \text{Age} + \text{Cholesterol_Level} + \text{Blood_Pressure_Systolic} + \text{Blood_Pressure_Diastolic} + \text{Smoking_Status} + \text{Obesity_Index} + \text{Diabetes_Status} + \text{Family_History_Heart_Disease}$

Os 3 principais nós da árvore de decisão (Modelo 3):

- $\text{Heart_Attack_Outcome} \sim \text{Age} + \text{Cholesterol_Level} + \text{Obesity_Index}$

AVALIAÇÃO DOS MODELOS

Os modelos 1 e 2 tiveram métricas muito similares, entretanto, como já apresentado, o modelo 2 tem um número muito menor de variáveis, o que aumenta a sua interpretação e simplicidade, algo imprescindível.

O modelo 3 apresenta métricas contraditórias, ruins para treino e ótimas para teste, portanto, não há confiabilidade.

Dito isso, o modelo escolhido foi o Modelo 2, que contém 8 variáveis e alto desempenho em todas as métricas:

- alta acurácia (87,7% para teste);
- excelente capacidade preditiva (AUC: 95,5% e GINI: 91,1%) ;
- boa separação de classes (KS: 75,3%).

CONCLUSÕES

Durante a análise realizada, identificamos que as principais causas associadas à incidência de ataque cardíaco nesta amostra incluem idade avançada, níveis elevados de colesterol, hipertensão, tabagismo, obesidade e histórico familiar de doenças cardíacas.

Desenvolvemos um modelo de regressão logística que apresentou bom desempenho e pode ser um aliado na identificação de perfis de pacientes que necessitam de mudanças de hábitos e implementação de estratégias de prevenção precoce.