Bootcamp: Analista de Dados com ênfase para Mercado Financeiro

Desafio

Módulo 2: Fundamentos em Ciências de Dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Criar virtualenv e realizar a instalação das bibliotecas necessárias para o trabalho.
2. Coleta de dados.
3. Analisar e realizar tratamento de dados.
4. Criar algoritmo de Machine Learning.
5. Avaliar os resultados
6. Conhecimento teórico ministrado nas videoaulas.

Enunciado

Você foi selecionado por um prestigiado hospital para desenvolver um modelo de classificação capaz de identificar pacientes com elevado risco de sofrer um ataque cardíaco. Como cientista de dados, você reconhece a importância crucial de detectar esses casos precocemente, visando proporcionar intervenções preventivas oportunas e salvar vidas. O hospital dispõe de um extenso conjunto de dados médicos, contendo informações de histórico médico, exames laboratoriais e registros de sintomas de diversos pacientes. Esses dados foram meticulosamente coletados ao longo de vários anos e

incluem informações sobre pacientes que tiveram ataques cardíacos confirmados, assim como aqueles que não tiveram.

Seu propósito é criar um modelo de classificação utilizando o algoritmo Random Forest, que seja competente para analisar essas informações e identificar padrões e características-chave indicativos de um elevado risco de ataque cardíaco. Isso viabilizará aos profissionais de saúde concentrarem seus esforços nos pacientes que mais necessitam de cuidados e intervenções preventivas.

Seu papel como cientista de dados é de extrema relevância, pois contribui para a saúde e bem-estar dos pacientes, disponibilizando aos profissionais médicos uma ferramenta valiosa para a detecção precoce e intervenção adequada em casos de elevado risco de ataques cardíacos.

Informações sobres os dados do dataset:

Atributo	Descrição	Valores/Significados
age	Idade	A idade do paciente em anos.
sex	Sexo	O sexo do paciente (0: feminino, 1: masculino).
cp	Chest Pain Type	Tipo de Dor no Peito
		- Value 1: typical angina
		- Value 2: atypical angina
		- Value 3: non-anginal pain
		- Value 4: asymptomatic
trtbps	Resting Blood Pressure	Pressão Arterial em Repouso - em mmHg.
chol	Cholesterol	Nível de colesterol do paciente em mg/dL.
fbs	Fasting Blood Sugar	Açúcar no Sangue em Jejum
		- 1: sim
		- 0: não
restecg	Resting Electrocardiographic Results	Resultados Eletrocardiográficos em Repouso
		- Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
		- Value 1: normal
		- Value 2: having ST-T wave abnormality (T wave inversions and/or ST)
thalachh	Maximum Heart Rate Achieved	Frequência Cardíaca Máxima Alcançada durante o teste de estresse.
exng	Exercise Induced Angina	Angina Induzida por Exercício
		- 1: sim
		- 0: não
oldpeak	ST Depression Induced by Exercise Relative to Rest	Alteração no segmento ST induzida pelo exercício em relação ao repouso.
slp	Slope of the Peak Exercise ST Segment	Inclinação do segmento ST de pico durante o exercício.
caa	Number of Major Vessels (0-3) Colored by Fluoroscopy	Número de principais vasos sanguíneos coloridos por fluoroscopia (0-3).
thall	Thalassemia	Tipo de Talassemia
		- 0: NULL (removido do conjunto de dados anteriormente)
		- Value 1: fixed defect (nenhum fluxo sanguíneo em alguma parte do coração)
		- Value 2: normal blood flow
		- Value 3: reversible defect (observado fluxo sanguíneo, mas não é normal)
output		Chance de ter um ataque cardíaco
		- 0: menos chance de ataque cardíaco
		- 1: mais chance de ataque cardíaco

Atividades

Para este desafio, os alunos deverão realizar as seguintes atividades:

1. Criar uma virtualenv e instalar as bibliotecas necessárias.
2. Coletar o arquivo heart.csv.
3. Analisar os dados coletados.
4. Avaliar a necessidade de tratamentos de dados ausentes.
5. Criação modelo de Random forest Classifier.
6. Avaliar os resultados obtidos.
7. Responder às questões teóricas e práticas do trabalho.

Dicas do professor:

1. Analise com cuidado os dados.
2. Realize todo tratamento de dados antes de responder às questões.
3. Antes de enviar as respostas, verifique se o gabarito está correto.
4. Atenção no momento de filtrar e corrigir dados (se necessário).
5. Tenha atenção no que pede cada questão.
6. Realize o balanceamento dos dados.
    a. RandomUnderSampler(random_state = 42)
    b. TomekLinks(sampling_strategy='all')
7. Separe 80% dos dados para treinar o algoritmo e 20% para testar.

a. Utilize o parâmetro de random_state=42.

8. Elimine dados duplicados, caso existam.

9. Siga fielmente todos os passos contidos no enunciado das questões.

10. O dataset utilizado no trabalho pode ser obtido no link:

    https://github.com/ProfLeandroLessa/classroom-datasets/tree/master/ANC/BTC%201/Desafio

11. Outras dicas de análises de dados e geração de gráficos em:

    https://leandrolessa.com.br/tutoriais/grafico-de-dispersao-como-criar-e-analisar-na-pratica/

12. Abaixo segue as versões utilizadas das bibliotecas neste trabalho:

```
VERSÕES DE BIBLIOTECAS UTILIZADAS
PANDAS: 2.0.1
SKLEARN: 1.2.2
```

ATENÇÃO PARA TRATAMENTO DE DADOS

Avalie se será necessário realizar tratamento de dados ausentes no datasets disponibilizado.

Instruções para correção de dados ausentes

1. Para dados numéricos, utilize a estratégia de exclusão dos dados.

2. Para os dados categóricos, utilize a moda dos valores da coluna.