**NAC3 - Aprendizado de maquina**

| **Nome dos Integrantes** | **RM** | **Turma** |
| --- | --- | --- |
| Integrante 1 Ricardo Modesto | RM | 86468 |
| Integrante 2 | RM |  |

**Data de entrega final**

Será considerado para avaliação e correção o ultimo commit do \*\* dia 11/09 (domingo) \*\*

O grupo deve responder ao Google forms : <https://forms.gle/EEYUkZiWcJ6Xduah6>

**Rubrica**

Os critérios de avaliação do projeto será conforme a rubrica.

* R1: Exploração de Dados (Até 1 ponto)
* R2: Preparação de Dados (Até 2 ponto)
* R3: Desenvolvimento do Modelo (Até 1 ponto)
* R4: Treinamento e Teste do Modelo (Até 2 ponto)
* R5: Validação e Otimização do Modelo (Até 3 ponto)
* R6: Conclusões Finais (Até 1 ponto)
* R7: Deploy (Até 2 pontos \*extra)

Nota\_final = R1 + R2 + R3 + R4 + R5 + R6 + R7

Siga todas as instruções deste noteboook, mas não limite suas análises apenas ao conteúdo visto em aula.

Nas celulas **Seu Código aqui...** significa que é necessário realizar a implementação de algum código. Utilize quantos blocos julgar necessário para criar o código de resposta.

Nas celulas **Responda** significa que é nessário uma resposta clara e objetiva (pode ser por texto, gráfico, imagem...) para a pergunta.

**Respeite a rubrica e não responda fora de ordem ou sequência**

**Análise preditiva de doeça cardíaca em pacientes**

Vamos trabalhar com um CONJUNTO DE DADOS DE DOENÇA CARDÍACA afim de predizer a presença ou não de doenças do coração em pacientes.

[dataset: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive](https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive).

**1. Exploração de Dados**

Carregue o dataset e faça uma visualização das primeiras linhas (identificação básica apenas confirmar se a base foi carregada corretamente).

In [1]:

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

**import** pandas **as** pd

**import** numpy **as** np

**%matplotlib** inline

In [2]:

*#Seu Código aqui...*

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

# Lê e carrega o arquivo para a memória

df = pd.read\_csv('./heart\_statlog\_cleveland\_hungary\_final.csv')

print(df)

**Pergunta:** Apresente (Escreva) uma descrição dos atributos (colunas) e seu significado:

Exemplo:

* id: Representa o fabricante
* fuel\_type: tipo de combustível (pode ser Gasolina ou Disel)
* price: Representa o valor do carro (**Variável Alvo, atributo que será predito pelo modelo de ML**)

**Responda**

Sua resposta aqui.......

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 1190 non-null int64

1 sex 1190 non-null int64

2 chest pain type 1190 non-null int64

3 resting bp s 1190 non-null int64

4 cholesterol 1190 non-null int64

5 fasting blood sugar 1190 non-null int64

6 resting ecg 1190 non-null int64

7 max heart rate 1190 non-null int64

8 exercise angina 1190 non-null int64

9 oldpeak 1190 non-null float64

10 ST slope 1190 non-null int64

11 target 1190 non-null int64

Table

Description automatically generated

**Análise profunda dos dados**

Faça uma análise exploratória dos dados. Utilize-se de gráficos e tabelas auxiliares, faça correlações entre os atributos para compreeder a natureza do problema.

In [3]:

*#Seu Código aqui...*

*Código 1:*

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

# Lê e carrega o arquivo para a memória

df = pd.read\_csv('./heart\_statlog\_cleveland\_hungary\_final.csv')

df.describe()

*Código 2:*

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

# Lê e carrega o arquivo para a memória

df = pd.read\_csv('./heart\_statlog\_cleveland\_hungary\_final.csv')

#df.describe()

df.hist(bins=100, figsize=(15, 15))

plt.show()

**Perguntas:**

1. Qual a relação entre a idade e colesterol?

R. Quanto maior a idade, o colesterol tende a aumentar.

1. O indice açucar no sangue > 120 mg/dl é mais frequênte em homens ou mulheres?

R. Em homens.

Graphical user interface, text, application, email

Description automatically generated

1. Qual a distribuição de idade do dataset?

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

# Lê e carrega o arquivo para a memória

df = pd.read\_csv('./heart\_statlog\_cleveland\_hungary\_final.csv')

df.groupby(['sex']).agg(['max','min','mean','median'])

1. Pessoas com presença de doença cardiaca são em sua maioria do grupo masculino ou feminino?

R. Masculino

A screenshot of a computer

Description automatically generated with medium confidence

**2. Preparação dos Dados**

**Lembre-se:** Trate esses dados dentro do contexto do problema. Justifique suas escolhas.

**Análise de valores inválidos**

Faça a análise de dados ausente, dados nulos e dados iguais a 0(zero).

In [5]:

*#Seu Código aqui...*

Text

Description automatically generated

**Pergunta:** Foram localizados dados ausente, dados nulos e dados iguais a 0(zero)? Se existe, como foram tratados (removidos ou mantidos ou substituidos ou ....)?

**Resposta:**

R. Não foram encontrados dados nulos e os valores zerados fazem parte da escala de dados.

Table

Description automatically generated

**Análise de Outliers**

Faça a análise de outliers.

In [ ]:

*#Seu Código aqui...*

*Análise*

*Text

Description automatically generated*

*Chart

Description automatically generated*

*Remoção:*

*Graphical user interface, application

Description automatically generated*

*Chart, line chart

Description automatically generated*

**Pergunta:** Foram localizados outliers? Se existe, como os dados foram tratados (removidos ou mantidos)?

**Resposta:**

R. Sim, foram encontrados outliers nas colunas numéricas, plotei para cada coluna numérica e outro contextualizando o dataSet. Outliers foram removidos.

**Tratamento de dados Paramétricos**

Faça o tratamento dos atributos categoricos, se existir. Manipule de forma conveniente para ser interpretado como um dado numérico.

In [9]:

*#Seu Código aqui...*

Graphical user interface

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Pergunta:** Quais são os atributos categoricos? Qual foi a técnica de preprocessamento utilizada?

**Resposta:**

R. Foram os atributos númericos: resting\_blood\_pressure cholesterol, max\_heart\_rate e oldpeak. Técnica utiliza Intervalo interquartil.

**Normalização dos dados**

Faça a manipulação adequada para normalizar os dados, se necessário.

In [ ]:

*#Seu Código aqui...*

Graphical user interface, text, application, email

Description automatically generated

**Pergunta:** Os dados foram normalizados? Qual foi a técnica de preprocessamento utilizada?

**Resposta:**

Sua resposta aqui.....

R. Sim, Scikit-Learn.

**3. Desenvolvimento do Modelo**

**Pergunta:** O conjunto de dados indica a aplicação de qual modelo de aprendizagem de maquina?

**Resposta:**

[ ] Aprendizado por reforço

[X] Aprendizado supervisionado

[ ] Aprendizado não supervisionado

**Pergunta:** Qual a técnica sugerida para este projeto?

**Resposta:**

[X] Regressão

[ ] Classificação

[ ] Clusterização

**Pergunta:** Escolha 3 modelos de algoritmos de ML para ser testado.

**Resposta:**

1. algoritimo ...**\_\_\_** Regressão Linear
2. algoritimo ...**\_\_\_** Análise Discriminante Linear
3. algoritimo ...**\_\_\_** Árvores de classificação e regressão

**Pergunta:** Explique como cada algoritmo funciona.

**Resposta:**

1. algoritimo ...**\_\_\_** Regressão Linear

Sua origem vem da correlação linear, que é a verificação da existência de um relacionamento entre duas variáveis. Ou seja, dado X e Y, quanto que X explica Y. Para isso, a regressão linear utiliza os pontos de dados para encontrar a melhor linha de ajuste para modelar essa relação.

1. algoritimo ...**\_\_\_** Análise Discriminante Linear

Regressão logística é um algoritmo de classificação tradicionalmente limitado a apenas problemas de classificação de duas classes. Se você tiver mais de duas classes, o algoritmo de Análise Discriminante Linear (LDA, na sigla em inglês) é a técnica de classificação linear preferida.

1. algoritimo ...**\_\_\_** Árvores de classificação e regressão

A representação do modelo da árvore de decisão é uma árvore binária. Cada nó representa uma única variável de entrada (x) e um ponto de divisão nessa variável (assumindo que a variável seja numérica).

**Dataset dos dados tratados**

Crie um subset dos dados apenas com os atributos que serão utilizados e exiba as primeiras linhas.

In [2]:

*#Seu Código aqui...*

Table

Description automatically generated

**Definição de X e Y**

Separe os subsets dos atributos em Y e X :

* Y = atributo dependente (alvo)
* X = Atributo independente (entradas)

In [ ]:

*#Seu Código aqui...*

Graphical user interface, text, application, email

Description automatically generated

**Divisão em treino e teste.**

Faça o split do dados para treinamento e teste

In [2]:

*#Seu Código aqui...*

Graphical user interface, text, application, email

Description automatically generated

**4. Treinamento e teste do modelo**

Defina a métrica de performance para avaliação dos modelos de ML

**Resposta:**

Sua resposta aqui.......

**Treinamento e teste**

Faça o treinamento e teste de dos 3 algoritmos escolhidos, aplique no treinamento a validação cruzada para um kfold de 5.

**Algoritmo 1**

In [18]:

*#Seu Código aqui...*

Text

Description automatically generated with medium confidence

**Algoritmo 2**

In [10]:

*#Seu Código aqui...*

**Algoritmo 3**

In [17]:

*#Seu Código aqui...*

Sumarize na tabela abaixo os resultados obtidos dos modelos de ML para o conjuntos de dados de treino e de testes

**Resposta:**

|  | **algoritmo1** | **algoritmo2** | **algoritmo3** |
| --- | --- | --- | --- |
| treino | X | X | X |
| teste | X | X | X |

**Pergunta:** Qual foi a performance de cada algoritmo? Explique os resultados e detalhe como a característica de cada algoritmo influenciou no resultado.

**Resposta:**

**5. Validação e Otimização do Modelo**

Faça a otimização dos modelos, tunning dos hiperparametros, utilize uma das técnicas: **RandomSearchCV** ou **GridSearchCV**.

Após a otimização dos parametros, treine novamente os 3 algoritmos de ML utilizando os parametos de otimização e compares seus resultados.

**Pergunta:** Qual a técnica escolhida? Por que?

**Resposta:**

**Algoritmo 1**

In [18]:

*#Seu Código aqui...*

**Algoritmo 2**

In [10]:

*#Seu Código aqui...*

**Algoritmo 3**

In [17]:

*#Seu Código aqui...*

Sumarize na tabela abaixo os resultados obtidos dos modelos de ML para o conjuntos de dados de treino e de testes

**Resposta:**

|  | **algoritmo1** | **algoritmo2** | **algoritmo3** |
| --- | --- | --- | --- |
| treino | X | X | X |
| teste | X | X | X |

**Pergunta:** Qual dos algoritmos obteve os melhores resultados? Quais caracaterísticas deste algoritmo podem ajudar a justificar tal resultado?

**Resposta:**

**Salvando o modelo desenvolvido**

Configure o modelo selecionado com os parâmetros otimizados obtidos anteriormente e salve

*# exemplo de como salvar o modelo "model" no arquivo "modelsave.pkl"*

*# import lib*

**import** pickle

*# salvar*

**with** open('modelsave.pkl', 'wb') **as** file:

pickle**.**dump(model, file)

In [26]:

*#Seu Código aqui...*

**6. Conclusões finais**

Ufa!!... Chegamos ao final deste projeto. Parabéns! De forma simplificada e completa passamos por todas as etapas de aprendizado de maquina. Outras etapas podem ser adicionadas no tratamento dos dados e desenvolvimento do modelo.

**Pergunta:**: O modelo desenvolvido obteve um resultado satisfatório? Justifique:

**Resposta**:

**Pergunta:**: O que faria diferente para melhorar resultado obtido? Justifique:

**Resposta**:

**Pergunta:**: As etapas de desenvolvimento se aplicam para um problema de classificação? Justifique:

**Resposta**:

**7. Deploy**

Etapa bonûs (Até 2 pontos \*extra)

Essa parte não é obrigatória, por esse motivo é considerada pontuação extra na nota para quem desenvolver.

Faça o deploy do modelo de ML desenvolvido em um webserver. Como sugestão utilize Flask ou FastAPI ou Django. Procure por referências na web.