Relatório Mineração de Dados

Base de dados: Horse Colic Data Set

Link: https://archive.ics.uci.edu/ml/datasets/Horse+Colic

Dados: Aspectos patológicos de cavalos que apresentam cólicas.

Número de Amostras: 300

Número de Atributos: 28

Tipos de atributos: numéricos e discretos.

Valores ausente: Sim

Passos:

1. Nomeação de todos atributos.
2. Especificação dos atributos que serão utilizados.
3. Identificação de atributos numéricos e discretos.
4. Especificação de atributos numéricos e discretos.
5. Identificação de dados faltantes.
6. Preenchimento de dados faltante de acordo com sua categoria. Obs.: Numéricos (Média); Discretos (Moda).
7. Geração de um arquivo com os dados ausentes preenchidos.
8. Impressão dos dados.

Código:

```
import pandas as pd


def main():

    collumns = ['Cirurgia', 'Idade', 'ID', 'Temperatura Retal', 'Pulso', 'Ritmo Respiratório', 'Temperatura das Extremidades', 'Pulso Periférico', 'Mucosas', 'Tempo de Preenchimento Capilar', 'Dor', 'Movimento Peristáltico',
                'Distensão Abdominal', 'Tubo Nasogratrico', 'Refluxo Nasogástrico', 'Ph do Refluxo Nasosgástrico', 'Exame Retal - Fezes', 'Abdomen', 'Hematócrito', 'Proteína Total', 'Aparência Paracentese', 'Proteína Paracentese', 'Resultado', 'Lesão Cirúrgica', 'Tipo da Lesão', 'Tipo da Lesão 2', 'Tipo da Lesão 3', 'Dados Patológicos']  # Todas as colunas
```

```
    useCollums = ['Cirurgia', 'Idade', 'Temperatura Retal', 'Pulso', 'Ritmo Respiratório', 'Temperatura das Extremidades', 'Pulso Periférico', 'Mucosas', 'Tempo de Preenchimento Capilar', 'Dor', 'Movimento Peristáltico',
                 'Distensão Abdominal', 'Tubo Nasogratrico', 'Refluxo Nasogástrico', 'Ph do Refluxo Nasosgástrico', 'Exame Retal - Fezes', 'Abdomen', 'Hematócrito', 'Proteína Total', 'Aparência Paracentese', 'Proteína Paracentese', 'Resultado', 'Lesão Cirúrgica', 'Dados Patológicos']  # Colunas que serão utilizadas

  continuousData = ['Temperatura Retal', 'Tempo de Preenchimento Capilar', 'Hematócrito',
                    'Ph do Refluxo Nasosgástrico', 'Proteína Total', 'Ritmo Respiratório', 'Proteína Paracentese', 'Pulso']  # Colunas com dados numéricos contínuos

  input_file = '../Dataset/horse-colic.data'  # Importação dos Dados
  df = pd.read_csv(input_file, delim_whitespace=True,
           names=collumns, usecols=useCollums, na_values='?')

  for campo in useCollums:  # Percorre a lista das colunas que estão sendo utilizadas
    if campo in continuousData:  # Checa se o campo atual pertence ao vetor de dados contínuos, se sim utilizada a média para preencher os valores nulos
      method = 'mean'
    else:  # Caso o dado não pertença ao conjuntos de dados contínuos será utilizado a moda
      method = 'mode'

    if method == 'mean':
      # Substituindo valores ausentes pela média
      mean = round(df[campo].mean(), 1)
      df[campo].fillna(mean, inplace=True)
    else:
      # Substituindo valores ausentes pela moda
      mode = df[campo].mode()[0]
      df[campo].fillna(mode, inplace=True)

  # Gera um arquivo csv com os todos os dados preenchidos pelo algoritmo
  df.to_csv('dados.csv')

  print(df.info())
```

```
if __name__ == "__main__":
    main()
```