Relatório de Pré processamento - inspeção de datasets

Cláudia Patrícia Silva Pimentel

E-mails: cacaupimentel@hotmail.com (mailto:cacaupimentel@hotmail.com), cacaupimentel@gmail.com (mailto:cacaupimentel@gmail.com), claudiapimentel@aluno.uema.br (mailto:claudiapimentel@aluno.uema.br)

4. Resumo e inspeções dos atributos de Post-Operative Patient Data Set (https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient)

4.1 Informações sobre a Fonte

1. Autores: Sharon Summers, Linda Woolery, Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Henry Tirri, Art B. Owen, Glenn Fung, Sathyakama Sandilya e R. Bharat Rao;
2. Artigos: Program LERS_LB 2.5 as a tool for knowledge acquisition in nursing, The use of machine learning program LERS_LB 2.5 in knowledge acquisition for expert system development in nursing, Unsupervised Bayesian visualization of high-dimensional data, Tubular neighbors for regression and classification e Rule extraction from Linear Support Vector Machines;
3. Publicação: Proceedings of the 4th Int. Conference on Industrial & Engineering Applications of AI & Expert Systems, Computers in Nursing 9, Conference on Industrial & Engineering Applications of AI & Expert;
4. Ano: 1991-2000;
5. Objetivo da Pesquisa: posposta de métodos de redução de dados em framework de probabilidade de similaridades com algoritmos não supervisionadas com o intuito de descrever como produzir imagens visuais de alta dimensionalidade de dados utilizando dados conjunto de dados do paciente pós-operatório (Kontkanen, Lahtinen, Myllymäki, 2000);
6. Resultados da Pesquisa: apresentam vários exemplos de visualização numa rede de modelos baysiana (Kontkanen, Lahtinen, Myllymäki, 2000) e no de Woolery, et al. (1991) apresentou que LERS (LEM2): 48% de precisão.

4.2 Importando o Post-Operative Patient Data Set (https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient)

Foram instalados os pacotes do pandas, numpy e pandas_profiling, para exploraçao dos dados.

In [1]:

```
# Instalando o pacote pandas
%pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\fcalp\anaconda3\lib\site-p
ackages (1.1.3)
Requirement already satisfied: pytz>=2017.2 in c:\users\fcalp\anaconda3\lib
\site-packages (from pandas) (2020.1)Note: you may need to restart the kerne
l to use updated packages.

Requirement already satisfied: numpy>=1.15.4 in c:\users\fcalp\anaconda3\lib
\site-packages (from pandas) (1.19.2)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\fcalp\anac
onda3\lib\site-packages (from pandas) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\fcalp\anaconda3\lib\site
-packages (from python-dateutil>=2.7.3->pandas) (1.15.0)
```

In [2]:

```
# Instalando o pacote numpy
%pip install numpy
```

```
Requirement already satisfied: numpy in c:\users\fcalp\anaconda3\lib\site-pa
ckages (1.19.2)Note: you may need to restart the kernel to use updated packa
ges.
```

In [3]:

```
# Instalando o pacote pandas_profiling
%pip install pandas_profiling
```

```
\lib\site-packages (from pandas_profiling) (2.11.2)
Requirement already satisfied: missingno>=0.4.2 in c:\users\fcalp\anaconda
3\lib\site-packages (from pandas_profiling) (0.4.2)
Requirement already satisfied: numpy>=1.16.0 in c:\users\fcalp\anaconda3\l
ib\site-packages (from pandas_profiling) (1.19.2)
Requirement already satisfied: confuse>=1.0.0 in c:\users\fcalp\anaconda3
\lib\site-packages (from pandas_profiling) (1.4.0)
Requirement already satisfied: htmlmin>=0.1.12 in c:\users\fcalp\anaconda3
\lib\site-packages (from pandas_profiling) (0.1.12)
Requirement already satisfied: matplotlib>=3.2.0 in c:\users\fcalp\anacond
a3\lib\site-packages (from pandas_profiling) (3.3.2)
Requirement already satisfied: visions[type_image_path]==0.6.0 in c:\users
\fcalp\anaconda3\lib\site-packages (from pandas_profiling) (0.6.0)
Requirement already satisfied: requests>=2.24.0 in c:\users\fcalp\anaconda
3\lib\site-packages (from pandas_profiling) (2.24.0)
Requirement already satisfied: pandas!=1.0.0,!=1.0.1,!=1.0.2,!=1.1.0,>=0.2
5.3 in c:\users\fcalp\anaconda3\lib\site-packages (from pandas_profiling)
(1.1.3)
Requirement already satisfied: ipywidgets>=7.5.1 in c:\users\fcalp\anacond
```

In [4]:

```
# Importando as bibliotecas necessárias
import pandas as pd
import numpy as np
from pandas_profiling import ProfileReport
```

4.3 Procedimentos de limpeza do arquivo

Na importação dos dados não houve nenhuma alteração ou ajuste de imediato para ser feito.

In [14]:

```
coding = 'utf-8',names=["L-CORE", "L-SURF", "L-O2", "L-BP", "SURF-STBL", "CORE-STBL", "BP-ST
```

Out[14]:

	L-CORE	L-SURF	L-O2	L-BP	SURF-STBL	CORE-STBL	BP-STBL	CONFORTO	decisão ADM-DECS
0	mid	low	excellent	mid	stable	stable	stable	15	A
1	mid	high	excellent	high	stable	stable	stable	10	S
2	high	low	excellent	high	stable	stable	mod-stable	10	A
3	mid	low	good	high	stable	unstable	mod-stable	15	A
4	mid	mid	excellent	high	stable	stable	stable	10	A
...	...	...	...	...	...	...	...	...	...
85	mid	mid	excellent	mid	unstable	stable	stable	10	A
86	mid	mid	excellent	mid	unstable	stable	stable	15	S
87	mid	mid	good	mid	unstable	stable	stable	15	A
88	mid	mid	excellent	mid	unstable	stable	stable	10	A
89	mid	mid	good	mid	unstable	stable	stable	15	S

90 rows × 9 columns

4.4 Inspeção e descrição dos Atributos

Woolery, et al. (1991) comenta que a tarefa de classificação do banco de dados em questão é determinar para onde os paciente, em área de recuperação pós-opertória, devem ser ecaminhados, destaca os que possuem hiptermia, por isso os atributos desta tabela fazem correspondência aproximada às mediçoes da temperatura corporal. Neste sentido as informações dos atributos diponibiliza de forma simplificada como a seguinte descrição:

1. L-CORE (temperatura interna do paciente em C):
   alto (> 37), médio (> = 36 e <= 37), baixo (<36)
2. L-SURF (temperatura da superfície do paciente em C):
   alto (> 36,5), médio (> = 36,5 e <= 35), baixo (<35)
3. L-O2 (saturação de oxigênio em%):
   excelente (> = 98), bom (> = 90 e <98),
   razoável (> = 80 e <90), ruim (<80)
4. L-BP (última medição da pressão arterial):
   alto (> 130/90), médio (<= 130/90 e> = 90/70), baixo (<90/70)
5. SURF-STBL (estabilidade da temperatura da superfície do paciente):
   estável, mod-estável, instável
6. CORE-STBL (estabilidade da temperatura central do paciente)
   estável, mod-estável, instável
7. BP-STBL (estabilidade da pressão arterial do paciente)
   estável, mod-estável, instável
8. CONFORTO (conforto percebido do paciente na alta, medido como um número inteiro entre 0 e 20)
9. decisão ADM-DECS (decisão de quitação):
   I (paciente encaminhado para Unidade de Terapia Intensiva),
   S (paciente preparado para ir para casa),
   A (paciente encaminhado para andar do hospital geral);

In [15]:

```
# Inspecionando os atributos do dataset
posoperatorio.describe()
```

Out[15]:

	L-CORE	L-SURF	L-O2	L-BP	SURF-STBL	CORE-STBL	BP-STBL	CONFORTO	decisão ADM-DECS
count	90	90	90	90	90	90	90	90	90
unique	3	3	2	3	2	3	3	5	4
top	mid	mid	good	mid	stable	stable	stable	10	A
freq	58	48	47	57	45	83	46	65	63

4.5 Analisando resultados

A seguir tens a análise completa pelo "ProfileReport".

In [16]:

```
profile = ProfileReport(posoperatorio, title='Pandas Profiling Report', explorative=True)
profile.to_notebook_iframe()
```

Summarize dataset: 22/22 [00:54<00:00, 2.48s/it, Completed]

100%

Generate report structure: 1/1 [00:37<00:00, 37.23s/it]

100%

Render HTML: 100% 1/1 [00:04<00:00, 4.10s/it]

Variable types

Categorical	9

Warnings

Dataset has 10 (11.1%) duplicate rows	**Duplicates**
`SURF-STBL` is uniformly distributed	**Uniform**

Reproduction

Analysis started	2021-04-13 00:48:21.897127
Analysis finished	2021-04-13 00:49:07.602652
Duration	45.71 seconds
Software version	pandas-profiling v2.11.0 (https://github.com/pandas-profiling/pandas-
Download configuration	config.yaml (data:text/plain;charset=utf-8,title%3A%20Pandas%20Pr %20%5B%27true%27%2C%20%27false%27%5D%0Adataset%3A%

Variables

L-CORE
Categorical

Distinct	3

Referências

Petri Kontkanen and Jussi Lahtinen and Petri Myllymäki and Henry Tirri. Unsupervised Bayesian visualization of high-dimensional data. KDD. 2000.

Art B. Owen. Tubular neighbors for regression and classification. Stanford University. 1999.

Glenn Fung and Sathyakama Sandilya and R. Bharat Rao. Rule extraction from Linear Support Vector Machines. Computer-Aided Diagnosis & Therapy, Siemens Medical Solutions, Inc.