UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METROPOLE DIGITAL
IMD1101 - Aprendizado de Máquina – 2024.2
Aula05 – Pré-processamento e Amostragem

Visando exercitar os conceitos aprendidos nas aulas de limpeza, transformação de dados e Redução de casos (amostragem), escolha uma base de dados que possua atributos numéricos (int ou float) e discretos (object) no repositório abaixo (https://www.dropbox.com/sh/f6i1fj8qpjpud9c/AABMYB-Yfc7jTOZHB_qHPdGBa?dl=0):

Figura 1. Repositório contendo bases de dados em CSV.

De posse da base, utilizando Python (Pandas) imprima as características da mesma (número de instâncias e atributos), conforme Figura 2.

```
1. Mostre as caracterísiticas dos atributos;
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 699 entries, 0 to 698
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Clump_Thickness       699 non-null    int64
 1   Cell_Size_Uniformity  699 non-null    int64
 2   Cell_Shape_Uniformity 699 non-null    int64
 3   Marginal_Adhesion     699 non-null    int64
 4   Single_Epi_Cell_Size  699 non-null    int64
 5   Bare_Nuclei           699 non-null    object
 6   Bland_Chromatin       699 non-null    int64
 7   Normal_Nucleoli       699 non-null    int64
 8   Mitoses               699 non-null    int64
 9   Class                 699 non-null    object
dtypes: int64(8), object(2)
memory usage: 54.7+ KB
None
```

```
2. Mostre a média dos atributos (excluindo o atributos classe);
Clump_Thickness         4.417740
Cell_Size_Uniformity    3.134478
Cell_Shape_Uniformity   3.207439
Marginal_Adhesion       2.806867
Single_Epi_Cell_Size    3.216023
Bland_Chromatin         3.437768
Normal_Nucleoli         2.866953
Mitoses                 1.589413
dtype: float64

3. Mostre a mediana dos atributos (excluindo o atributos classe);
Clump_Thickness         4.0
Cell_Size_Uniformity    1.0
Cell_Shape_Uniformity   1.0
Marginal_Adhesion       1.0
Single_Epi_Cell_Size    2.0
Bland_Chromatin         3.0
Normal_Nucleoli         1.0
Mitoses                 1.0
dtype: float64
```

Figura 2. Visualização das características da base – Python (Pandas).

A seguir, execute as seguintes etapas relacionas à limpeza e transformação dos atributos da base escolhida:

1. Verifique se há *missing values.* Se houver aplique um dos métodos relacionados a esse tipo de problema (mostrados nas aulas);
2. Aplique as transformações necessárias para os atributos numéricos e discretos;
3. Salve uma versão em CSV dessa base limpa e transformada.

De posse da versão limpa e transformada, crie diversas amostragens dos dados, de acordo com o que se pede abaixo:

1. Amostragem simples de 30% e sem reposição;
2. Amostragem simples de 30% e com reposição;
3. Amostragem simples de 50% e sem reposição;
4. Amostragem simples de 50% e com reposição;
5. Amostragem estratificada de 50% (mesmas proporções);
6. Amostragem simples de 70% e sem reposição;
7. Amostragem simples de 70% e com reposição;
8. Amostragem estratificada de 70% (mesmas proporções).

Por último, submeta, via SIGAA, o seu Jupyter Notebook que executa todos os passos utilizados para limpar e transformar a base de dados escolhida, além daqueles utilizados para criar as diversas amostragens exigidas. Sua submissão valerá a presença referente a essa aula assíncrona.

Bom trabalho!