Pré-processamento

Germano C. Vasconcelos
Centro de Informática - UFPE

Passos Importantes no Tratamento da Base de Dados Não-codificada

- Histogramas para análise de distribuição
- Box plots e análises estatísticas (min, max, median, mean, quartiles)
- Análise de dados ausentes (missing data)
- Análise de outliers
- Codificação de dados categóricos
  - Integer encoding (quando houver ordem)
  - One hot encoding (variável com poucos valores – exemplo 3 ou 5)
  - Binary encoding (váriavel com muitos valores)
- Normalização de dados numéricos

Outros Passos Relevantes

- Binning para definir melhores agrupamentos
- Análise de correlação das variáveis independentes
- Análise de correlação com a variável dependente
- Seleção de variáveis
- Combinação de váriáveis (soma, divisão)

Alguns Links Úteis

- https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/
- https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/
- Using statistics: How to understand population distributions?
- https://www.analyticsvidhya.com/blog/2014/07/statistics/

Códigos em Python

- MLP Titanic Kaggle
- https://colab.research.google.com/drive/13QQMPWrYi84BV8_dic7c-XC40CO1nwuD?authuser=1
- https://bit.ly/2YjGD7K
- 100 Days of ML Code
- https://github.com/Avik-Jain/100-Days-Of-ML-Code
- Histogramas e Binning
- https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/04.05-Histograms-and-Binnings.ipynb