Modelagem e Preparação de Dados para Aprendizado de Máquina

Professor Luis Enrique Zárate

Exemplo Prático e Desafio: Construção de um modelo de aprendizado para descrever o perfil dos indivíduos que sofrem com a doença do “Colesterol Alto”.

Objetivo: Aplicar os conceitos de modelagem e preparação de dados para descrição do perfil de pessoas que sofrem com “Colesterol Alto”.

Origem dos dados: Estudo da Base de Dados sobre Pesquisa Nacional de Saúde - PNS 2013: percepção do estado de saúde, estilos de vida e doenças crônicas.

- A base de dados original pode ser encontrada no seguinte endereço: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=microdados

- O dicionário de dados pode ser encontrado em: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=PNS/2013/Microdados/Documentacao

- Informações referente ao estudo está disponível na biblioteca nacional do IBGE, e pode ser acessado no seguinte endereço: https://biblioteca.ibge.gov.br/visualizacao/livros/liv94074.pdf

Detalhamento das bases de dados:

- Informações Sobre o Dataset original: Variáveis = 942, Registros = 205546

- A partir da base de dados original foi extraído um dataset contendo registros de pessoas diagnosticadas com “Colesterol Alto”, e para contrapor ao modelo a ser construído, foi inserida a mesma quantidade de registros de pessoas que não sofrem da doença. O total de registros é de 14599.

- O conjunto de dados selecionados pode ser encontrado no arquivo: Dados.xlsx (primeira aba)

- Um modelo conceitual, construído por conhecimento explícito e tácito pode ser encontrado no arquivo: Mapa Conceitual.jpg

Procedimentos a serem seguidos de acordo à metodologia PICTOREA

1- Selecionar as variáveis mais relevantes de acordo ao mapa conceitual previamente construído.

2- Realizar uma análise descritiva prévia acerca do número de casos (Colesterol alto) por estado brasileiro e regiões do Brasil. O objetivo é tomar decisão se o modelo vai ser construído a nível brasil, região ou estado.
3- Após aplicada a análise do ítem 2, aplicar estatística descritiva (médias, medianas, desvio padrão, moda, histogramas, etc) das variáveis que compõem o conjunto de dados "alvo" do estudo. O objetivo desta etapa é explorar o conjunto de dados.
4- Aplicar uma observação univariada para detectear registros com inconsistências. Tomar decisão acerca da eliminação ou não desses registros e/ou variáveis.
5- Aplicar uma análise acerca da presença de dados ausentes. Avaliar a possibilidade de fusão de variáveis de forma a diminuir a presença de dados ausentes.
6- Propor e aplicar estratégias univariada para análise de outliers. Eliminar registros ou variáveis contendo outliers.
7- Propor estratégias para avaliar inconsistências em registros. Avaliar a coerência nos domínios entre variáveis e registros.
8- Aplicar estratégias para seleção de atributos univariada. Sugere-se avaliar a força de cada variável para separação de classes, se o problema escolhido for de classificação.
9- Considerando a etapa 8, avaliar a possibilidade de discretizar a variável de forma a tornar mais evidente a efetividade da variável na tarefa classificatória.
10- Aplicar estratégias univariadas para seleção de atributos sobre dados numéricos utilizando análise de distanciamento entre distribuições de densidade.
11- Avaliar a posibilidade da discretização de variáveis para aplicação de algoritmos de classificação ou clusterização. É sugerido que na presença de dados mixtos se opte por discretizar os dados numéricos, obtendo dados categóricos.
12- Realizar uma avaliação final descrevendo as restrições adotadas durante o processo de preparação da base de dados. Avaliar a representatividade da base de dados. Colocar restrições ao modelo ser obtido.
13-Aplicar técnica de aprendizado de máquina.

Observações:

- A experiência prática mostra que novas estratégias e procedimentos podem ser aplicados para obtenção de um conjunto de dados relevante ao problema.

- O conjunto disponibilizado na planilha contêm 3 abas, sendo a última um resultado possível de ser atingido pelo processo de preparação de dados.