

Machine Learning com Python

Prof. Luciano Galdino

Os dados

Informações provenientes de contagens, medidas, observações e respostas, formando conjuntos de dados.

Conjuntos de Dados:

- 1) População: relação de todos os dados de interesse. São extraídos parâmetros.
- 2) Amostra: é um subconjunto da população. São extraídas estatísticas.

Classificação dos dados

- Qualitativos: atributos não numéricos.
 - Nominais: Denominações (cores, gênero, raça, títulos...)
 - Ordinais: atributos que podem ser classificados (Ex: classificação de filmes mais assistidos, grau de escolaridade, nível de satisfação...).
- Quantitativos: medidas numéricas ou de contagem.
 - Discreto: valores finitos ou enumeráveis (quantidade de pessoas numa sala, número de carros em um estacionamento...)
 - Contínuo: infinitos valores possíveis num intervalo (renda, tempo, altura...).

Planejamento do estudo Estatístico

- 1) Identificação da população e das variáveis.
- 2) Plano para coleta de dados.
- 3) Coleta de dados.
- 4) Descrição dos dados usando técnicas da estatística descritiva.
- 5) Interpretação dos dados e tomadas de decisões através da estatística inferencial.
- 6) Análise de possíveis erros.

Observações e experimentos

Observações: elementos analisados não foram sujeitos a interferências (Ex.: pesquisa eleitoral).

Experimentos: elementos foram manipulados para obtenção dos resultados (Ex.: análise de medicamentos em testes).

Amostragem

- Amostragem é a medição ou contagem de parte de uma população.
- Censo é a medição ou contagem de toda a população.
- Amostras aleatórias: Todos os elementos tem chances iguais de serem relacionados. Pode ser com reposição ou sem reposição.
- Amostras tendenciosas: não representam a população.

Técnicas de Amostragem

- 1) Aleatória Simples: Seleção executada por meio de sorteio, sem nenhum filtro.
- 2) Estratificada: Divisão da população em grupos e seleção aleatória de uma amostra de cada grupo. (Ex: divisão por região, classe social, religião...).
- 3) Conglomerado (Agrupamento): Divisão da população em grupos com características similares, porém heterogêneas, e seleção aleatória de alguns grupos para analisar todos os elementos destes grupos. (Ex.: Divisão da população de escolas estaduais por região, enfermeiros de uma rede de hospitais...).
- 4) Sistemática: Membros da população são ordenados numericamente e são selecionados aleatoriamente, obedecendo uma sequência numérica. (Ex.: criação de números para cada amostra e seleção obedecendo uma ordem numérica).