**Avaliação de Pré-processamento de Dados para Aprendizado de Máquina**

**1. Qual é o principal objetivo do pré-processamento de dados no aprendizado de máquina para engenharia de sistemas de processos?**  
a. Aumentar a complexidade computacional  
b. Melhorar a precisão e confiabilidade do modelo  
c. Gerar mais pontos de dados  
d. Criar ruído artificial

**2. Na fórmula do filtro Savitzky-Golay x(t<em>)<em>{suavizado} = \sum</em>{k=1}^m b\_k (t</em>)^k, o que b\_k representa?**  
a. Tamanho da janela  
b. Variável de tempo  
c. Coeficientes polinomiais  
d. Fator de ruído

**3. Qual das seguintes NÃO é uma categoria principal de métodos de seleção de variáveis?**  
a. Métodos de Filtro  
b. Métodos Wrapper  
c. Métodos Neurais  
d. Métodos Embutidos

**4. O identificador de Hampel para detecção de outliers usa qual estatística robusta?**  
a. Desvio padrão  
b. Média  
c. Moda  
d. Mediana

**5. Qual é a fórmula correta para Média Móvel Simples (SMA)?**  
a. x(t)<em>{suavizado} = \prod</em>{j=0}^{W-1} x(t-j)<em>{bruto} b. x(t)</em>{suavizado} = \sum\_{j=0}^{W-1} x(t-j)<em>{bruto} / W c. x(t)</em>{suavizado} = \max\_{j=0}^{W-1} x(t-j)<em>{bruto} d. x(t)</em>{suavizado} = \min\_{j=0}^{W-1} x(t-j)\_{bruto}

**6. Qual método é computacionalmente mais intensivo para seleção de variáveis?**  
a. Métodos Embutidos  
b. Métodos de Filtro  
c. Métodos Wrapper  
d. Métodos Estatísticos

**7. O cálculo da Distância de Mahalanobis assume o que sobre a distribuição dos dados?**  
a. Distribuição bimodal  
b. Distribuição uniforme  
c. Distribuição exponencial  
d. Distribuição Gaussiana

**8. Qual é a principal vantagem do filtro Savitzky-Golay sobre os filtros de média móvel?**  
a. Implementação mais simples  
b. Melhor preservação das características do sinal  
c. Tempo de processamento mais rápido  
d. Menor custo computacional

**9. Na detecção multivariada de outliers, o que significa MCD?**  
a. Detecção de Correlação Múltipla  
b. Determinação de Covariância Mínima  
c. Distância de Correlação Máxima  
d. Determinante de Covariância Mínima

**10. Qual técnica NÃO é tipicamente usada para lidar com dados ausentes?**  
a. Imputação hot-deck  
b. Imputação pela média  
c. Detecção de outliers  
d. Imputação KNN

**11. Qual é a principal limitação dos métodos univariados de detecção de outliers?**  
a. São muito intensivos computacionalmente  
b. Não podem lidar com dados numéricos  
c. Ignoram interações entre variáveis  
d. Requerem muita memória

**12. A regra 3-sigma para detecção de outliers marca pontos de dados além de:**  
a. \mu \pm 2\sigma  
b. \mu \pm 2.5\sigma  
c. \mu \pm 3\sigma  
d. \mu \pm 4\sigma

**13. Qual algoritmo é usado para detecção robusta de outliers no Sklearn?**  
a. FAST-MCD  
b. RANSAC  
c. K-means  
d. DBSCAN

**14. Qual é uma característica principal dos Métodos de Filtro para seleção de variáveis?**  
a. Requerem treinamento do modelo  
b. Usam medidas estatísticas  
c. São computacionalmente intensivos  
d. Sempre consideram interações entre variáveis

**15. Qual método é mais apropriado para detectar outliers em distribuições complexas e não-Gaussianas?**  
a. Regra 3-sigma  
b. Distância de Mahalanobis  
c. Métodos baseados em densidade  
d. Média móvel simples

**16. Qual é o principal objetivo do algoritmo RANSAC?**  
a. Imputação de dados  
b. Seleção de características  
c. Ajuste robusto de modelo  
d. Remoção de ruído

**17. Na detecção de outliers em séries temporais, o que é analisado para identificar potenciais outliers?**  
a. Pontos de dados brutos  
b. Resíduos do modelo  
c. Médias móveis  
d. Coeficientes de correlação

**18. Qual afirmação sobre detecção de outliers baseada em PCA está correta?**  
a. Não pode lidar com dados multivariados  
b. Requer distribuição Gaussiana  
c. Pode distinguir diferentes tipos de outliers  
d. Só funciona com dados categóricos

**19. MAD no identificador de Hampel significa:**  
a. Desvio Absoluto Máximo  
b. Desvio Absoluto da Mediana  
c. Diferença Aritmética Média  
d. Detecção de Anomalia Múltipla

**20. Qual é uma vantagem principal dos Métodos Embutidos para seleção de características?**  
a. São mais rápidos que métodos wrapper  
b. Sempre fornecem resultados ótimos  
c. Não requerem ajuste de parâmetros  
d. Funcionam apenas com modelos lineares

Gabarito (Sempre revise as respostas geradas por IA para verificar a precisão)

1. b
2. c
3. c
4. d
5. b
6. c
7. d
8. b
9. d
10. c
11. c
12. c
13. a
14. b
15. c
16. c
17. b
18. c
19. b
20. a