Sistemas Inteligentes Aplicados - Trabalho 1

Introdução ao KDD e pré-processamento de dados

Abril de 2024

Aluno: Augusto Azevedo / RA: 2420660

Introdução

O trabalho tem como objetivo desenvolver e avaliar modelos de aprendizado de máquina para a predição de doenças cardiovasculares (DCV). As DCV são as principais causas de morte globalmente, portanto, identificar eficazmente os riscos é crucial para intervenções preventivas e tratamentos eficazes.

Descrição do Dataset

O dataset utilizado neste projeto contém 918 registros de pacientes, cada um com 12 atributos relacionados à saúde cardiovascular. As variáveis incluem idade, sexo, tipo de dor no peito, pressão arterial em repouso, colesterol, glicemia em jejum, resultados do eletrocardiograma em repouso, frequência cardíaca máxima alcançada, angina induzida por exercício, depressão ST induzida por exercício em relação ao repouso, a inclinação do segmento ST do pico do exercício e a presença ou ausência de doença cardíaca (variável alvo).

Transformações Aplicadas nos Dados

As transformações aplicadas variam conforme o modelo de aprendizado de máquina utilizado:

- **Codificação de Variáveis Categóricas:** Utilizei o One-Hot Encoding para converter variáveis categóricas como 'Sex', 'ChestPainType', 'ExerciseAngina', 'ST_Slope', e 'RestingECG' em formatos numéricos. Esta transformação é crucial pois muitos algoritmos de aprendizado de máquina não podem processar diretamente dados categóricos. A codificação permite que os modelos interpretem corretamente.

- **Normalização de Variáveis Numéricas:** Utilizei o StandardScaler para normalizar variáveis como 'Age', 'RestingBP', 'Cholesterol', 'MaxHR', e 'Oldpeak'. Esta etapa é necessária para modelos como SVM e k-NN, que são sensíveis à escala das features, garantindo que todas as variáveis contribuam igualmente para o resultado.

Modelos Avaliados

Três modelos de aprendizado de máquina foram selecionados e avaliados:

1. **Árvore de Decisão (Decision Tree):** Modelo simples e altamente interpretável. Foi configurada com uma profundidade máxima de 5 para evitar overfitting.
2. **Support Vector Machine (SVM):** Modelo robusto a espaços de alta dimensão, usando um kernel linear e um parâmetro de regularização C de 1.0.
3. **k-Nearest Neighbors (k-NN):** Baseado em instância, este modelo considerou os 5 vizinhos mais próximos para a classificação.

Parâmetros Alterados

Para cada modelo, ajustaram-se parâmetros específicos:

- **Árvore de Decisão:** max_depth ajustado para evitar overfitting, com valores testados de 5 e 7.
- **SVM:** C, o parâmetro de regularização, com valores de 1.0 e 0.5, para ver o efeito na margem de decisão e na tolerância a erros.
- **k-NN:** n_neighbors, ajustado para 5 e 7, para explorar como a quantidade de vizinhos afeta a classificação.

Treinamento e Avaliação

A performance foi avaliada com base em métricas como precisão, recall e f1-score. Resultados detalhados para cada modelo incluem:

- **Árvore de Decisão:**
    - Precisão: 86%
    - F1-Score: 0.85
    - Matriz de Confusão: 64 TN, 13 FP, 13 FN, 94 TP.
- **SVM:**
    - Precisão: 85%
    - F1-Score: 0.85
    - Matriz de Confusão: 67 TN, 10 FP, 17 FN, 90 TP.

- **k-NN:**
    - Precisão: 83%
    - F1-Score: 0.82
    - Matriz de Confusão: 61 TN, 16 FP, 16 FN, 91 TP.

Definição do Melhor Resultado

O modelo de Árvore de Decisão mostrou-se ligeiramente superior, com a maior precisão geral de 86%. Este modelo também ofereceu um bom equilíbrio entre precisão e recall, tornando-o o mais eficaz para o dataset utilizado.

Conclusão

Esta documentação resume as etapas seguidas, as transformações aplicadas, os resultados obtidos e os parâmetros ajustados nos modelos de aprendizado de máquina utilizados para prever doenças cardiovasculares. Utilizei dados na internet, vídeo aulas e inteligência artificial para realização deste trabalho.