Sistemas Inteligentes Aplicados - Trabalho 1

Introdução ao KDD e pré-processamento de dados

Maio de 2024

Aluno: Augusto Azevedo / RA: 2420660

Introdução

Este projeto é dedicado ao desenvolvimento e avaliação de modelos de aprendizado de máquina para a predição de doenças cardiovasculares (DCV), a principal causa de morte em todo o mundo. As DCV são responsáveis por aproximadamente 17,9 milhões de mortes anuais, representando cerca de 31% de todas as mortes globais. A identificação eficaz dos riscos associados às DCV é crucial para implementar intervenções preventivas e otimizar tratamentos. Além de abordar um problema crítico de saúde pública, este trabalho explora como técnicas avançadas de análise de dados podem ser aplicadas para melhorar os resultados de saúde.

Descrição do Dataset

O dataset utilizado neste estudo contém 918 registros de pacientes, caracterizados por 12 atributos essenciais para o diagnóstico e prognóstico de DCV. 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. A compreensão desses dados é fundamental para identificar padrões e correlações indicativas da presença de doenças cardiovasculares.

Informações de Atributo:

1. Idade: idade do paciente [anos]
2. Sexo: sexo do paciente [M: Masculino, F: Feminino]
3. ChestPainType: tipo de dor torácica [TA: Angina típica, ATA: Angina atípica, NAP: Dor não anginosa, ASY: Assintomática]
4. PA em repouso: pressão arterial em repouso [mm Hg]
5. Colesterol: colesterol sérico [mm/dl]

6. BS em jejum: glicemia em jejum [1: se BS em jejum > 120 mg/dl, 0: caso contrário]
7. ECG em repouso: resultados do eletrocardiograma em repouso [Normal: Normal, ST: com anormalidade das ondas ST-T (inversões das ondas T e/ou elevação ou depressão de ST > 0,05 mV), HVE: mostrando hipertrofia ventricular esquerda provável ou definitiva pelos critérios de Estes]
8. MaxHR: frequência cardíaca máxima alcançada [valor numérico entre 60 e 202]
9. ExercícioAngina: angina induzida por exercício [Y: Sim, N: Não] Oldpeak: oldpeak = ST [Valor numérico medido na depressão]
10. ST_Slope: a inclinação do segmento ST do pico do exercício [Up: subida, Flat: flat, Down: downsloping]
11. HeartDisease: classe de saída [1: doença cardíaca, 0: Normal]

Este conjunto de dados foi criado combinando diferentes conjuntos de dados já disponíveis de forma independente, mas não combinados anteriormente. Neste conjunto de dados, 5 conjuntos de dados cardíacos são combinados em 11 características comuns, o que o torna o maior conjunto de dados de doenças cardíacas disponível até agora para fins de pesquisa. Os cinco conjuntos de dados usados para sua curadoria são:

Cleveland: 303 observações

Húngaro: 294 observações

Suíça: 123 observações

Long Beach VA: 200 observações

Conjunto de dados Stalog (coração): 270 observações

Total: 1190 observações

Duplicado: 272 observações

Conjunto de dados final: 918 observações

Transformações Aplicadas nos Dados

Transformações dos dados foram meticulosamente planejadas para facilitar a análise por diversos algoritmos de aprendizado de máquina:

- **Codificação de Variáveis Categóricas:** Empregamos o One-Hot Encoding para transformar variáveis categóricas como 'Sex', 'ChestPainType', 'ExerciseAngina', 'ST_Slope' e 'RestingECG' em formatos numéricos. Isso é essencial para que os modelos de aprendizado de máquina possam interpretar as informações sem viés.
- **Normalização de Variáveis Numéricas:** Utilizamos o StandardScaler para ajustar as variáveis numéricas de modo que tenham média zero e desvio padrão unitário. Isso é crucial para modelos como SVM e k-NN, garantindo que as variáveis contribuam equitativamente para o resultado da análise.

Modelos Avaliados

Exploramos três modelos de aprendizado de máquina distintos para avaliar sua eficácia na previsão de DCV:

- **Árvore de Decisão (Decision Tree):** Este modelo oferece alta interpretabilidade e foi limitado a uma profundidade máxima de 5 para evitar overfitting.
- **Support Vector Machine (SVM):** Com sua eficácia em espaços de alta dimensão, o SVM foi configurado com um kernel linear e um parâmetro de regularização C ajustado para explorar diferentes tolerâncias a erros.
- **k-Nearest Neighbors (k-NN):** Este modelo intuitivo baseia-se na proximidade das características para classificar novos casos, com ajustes no número de vizinhos para otimizar a classificação.

Parâmetros Alterados

Os parâmetros de cada modelo foram cuidadosamente ajustados para melhorar o desempenho e compreender seu impacto nas previsões:

- **Árvore de Decisão:** Testamos max_depth de 5 e 7.
- **SVM:** O parâmetro C foi explorado com valores de 1.0 e 0.5.
- **k-NN:** Variamos n_neighbors entre 5 e 7.

Resultados do código:

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 918 entries, 0 to 917

Data columns (total 12 columns):

#	Column	Non-Null	Count Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	918 non-null	int64
4	Cholesterol	918 non-null	int64
5	FastingBS	918 non-null	int64
6	RestingECG	918 non-null	object
7	MaxHR	918 non-null	int64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	918 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	int64

dtypes: float64(1), int64(6), object(5)

memory usage: 86.2+ KB

None

Age 0

Sex 0

ChestPainType	0
RestingBP	0
Cholesterol	0
FastingBS	0
RestingECG	0
MaxHR	0
ExerciseAngina	0
Oldpeak	0
ST_Slope	0
HeartDisease	0

```
dtype: int64
              precision    recall  f1-score   support

           0       0.83      0.83      0.83        77
           1       0.88      0.88      0.88       107

    accuracy                           0.86       184
   macro avg       0.85      0.85      0.85       184
weighted avg       0.86      0.86      0.86       184

[[64 13]
 [13 94]]
              precision    recall  f1-score   support

           0       0.80      0.87      0.83        77
           1       0.90      0.84      0.87       107

    accuracy                           0.85       184
   macro avg       0.85      0.86      0.85       184
weighted avg       0.86      0.85      0.85       184

[[67 10]
 [17 90]]
              precision    recall  f1-score   support

           0       0.79      0.79      0.79        77
           1       0.85      0.85      0.85       107

    accuracy                           0.83       184
   macro avg       0.82      0.82      0.82       184
weighted avg       0.83      0.83      0.83       184

[[61 16]
 [16 91]]
```

Treinamento e Avaliação

Avaliamos os modelos com base em precisão, recall e f1-score, com a Árvore de Decisão alcançando a melhor precisão de 86%, demonstrando um equilíbrio efetivo entre sensibilidade e especificidade.

Conclusão

Estudo sobre técnicas de aprendizado de máquina na previsão de doenças cardiovasculares, utilizando um conjunto de dados detalhados e transformações de dados bem planejadas. A análise realizada pode servir como base para futuras investigações que poderiam expandir a aplicabilidade dos modelos em contextos clínicos reais. Espera-se que a continuação deste trabalho inclua a exploração de mais variáveis e a aplicação de modelos computacionais ainda mais sofisticados, como redes neurais profundas.

Referências

- Organização Mundial da Saúde. (2021). Cardiovascular diseases (CVDs).
- Scikit-learn Documentation.
- Pandas Documentation.