Comparação de modelos preditivos de doenças cardíacas usando técnicas de classificação de aprendizado de máquina

Aprendizado de Máquina(2023-2) – Profª Flávia Bernardini
Italo Leite Ferreira Portinho – aluno de Mestrado do PGC da UFF

Sumário

Doenças Cardiovasculares

- Desordens do coração e dos vasos sanguíneos
- Maior causa de morte no mundo, segundo a OMS
- Doenças cardiovasculares no Brasil
- Fatores de risco
- Identificação dos sintomas e tratamento adequado
- Aprendizado de máquina e doenças cardíacas

Revisão da literatura

- Eldouh et al.(2023)
- Neutrosophic AHP(analytical hierarchy process)
- Pesos para cada atributo e regras de associação
- Matriz de comparação
- Atributos de maior suporte são usados
- Random Forest, Bagging & DTree

Revisão da literatura

Table 3. Comparison matrix between 13 features.

Column name in dataset	Target class	antecedent support	consequent support	Support	confidence	lift	leverage	Conviction
Age	0	0.8537	0.9756	0.8293	0.9714	0.9957	-0.0036	0.8537
	1	0.9756	0.8537	0.8293	0.8500	0.9957	-0.0036	0.9756
Sex	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
CP	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
trestbps	0	0.7755	0.7959	0.5714	0.7368	0.9258	-0.0458	0.7755
	1	0.7959	0.7755	0.5714	0.7179	0.9258	-0.0458	0.7959
fbs	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
restecg	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
thalach	0	0.7363	0.7802	0.5165	0.7015	0.8991	-0.058	0.7363
	1	0.7802	0.7363	0.5165	0.6620	0.8991	-0.058	0.7802
exang	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
oldpeak	0	0.650	0.875	0.525	0.8077	0.9231	-0.0437	0.650
	1	0.875	0.650	0.525	0.6000	0.9231	-0.0437	0.875
slope	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
ca	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf
thal	0	1.0	1.0	1.0	1.0	1.0	0.0	Inf
	1	1.0	1.0	1.0	1.0	1.0	0.0	Inf

El-Douh et al., Heart Disease Prediction under Machine Learning and Association Rules under Neutrosophic Environment

Revisão da literatura

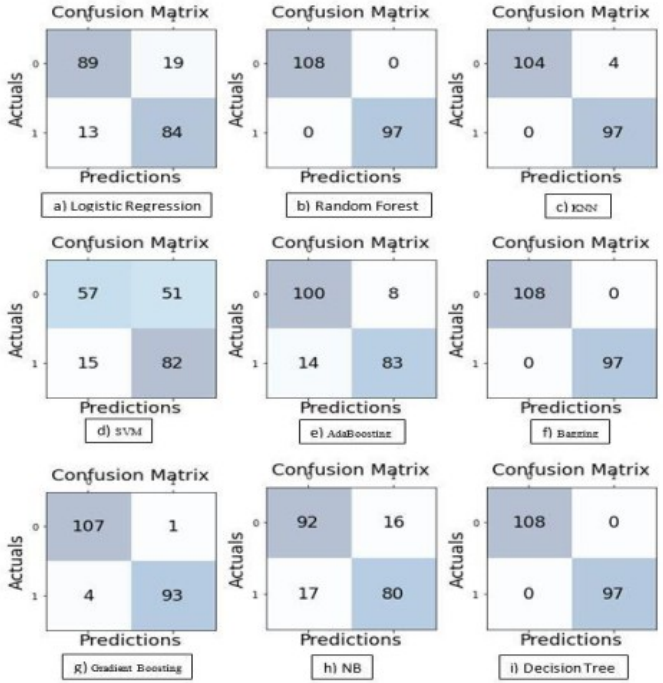

Figure 6. The confusion matrices.

Revisão da literatura

- Sabay et al. (2018)
- Datasets médicos carecem de poucos registros;
- Preocupação com dados sensíveis e pessoais
- Dados sintéticos
- Linguagem R e o pacote SynthPop
- 60.000 registros no dataset

O Dataset

- Dataset Cleveland 14, disponibilizado pela UCI
- Dataset original possui 76 atributos
- Porção relevante
- Muito usado em outros estudos
- Descrição dos atributos
- 303 registros, 14 atributos

O Dataset

Tabela 1. Descrição dos atributos do dataset.

Atributo	Descrição
age	Idade em anos
sex	gênero (1 = masculino; 0 = feminino)
cp	Dor torácica (0 = angina típica; 1 = angina atípica; 2 = não anginosa; 3 = assintomático)
trtbps	Pressão arterial em repouso (mm Hg)
chol	Colesterol Total
fbs	Glicemia em jejum > 120 mg/dl (1 = verdadeiro; 0 = falso)
restecg	Eletrocardiograma(0 = normal; 1 = ST-T anormal; 2 = hipertrofia do ventrículo esquerdo)
thalachh	Frequência cardíaca máxima
exng	Angina induzida por atividade física (1 = sim; 0 = não)
oldpeak	Depressão do ST induzida por atividade física em relação ao repouso
slp	Inclinação do segmento ST durante exercício (0 = aclive; 1 = linear; 2 = declive)
caa	Número de vasos coloridos na fluoroscopia (0 – 3)
thall	0 = normal; 1 = dano irreversível; 3 = dano reversível
output	Diagnóstico (0 = normal; 1 = chance aumentada para doença cardíaca)

O Dataset

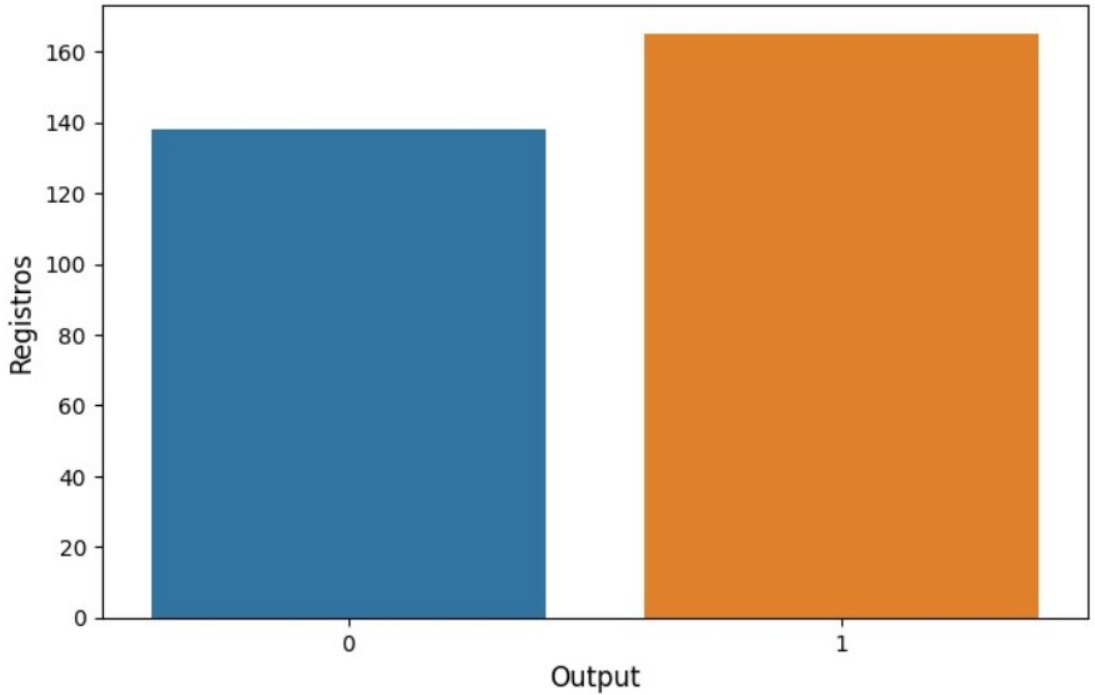

Figura 1. Distribuição dos valores nas classes do atributo alvo

O Dataset

```
# Confirma inexistência de atributos não preenchidos
dataset.isnull().sum()
```

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

O Experimento

- Colab, Jupyter, Python 3, SkLearn, TensorFlow
- KNN, SVM, Dtree, Random Forest, Naive Bayes & RNA
- Holdout e Validação Cruzada(0.47%)
- Acurácia, matriz de confusão, f1 score e curva ROC

O Experimento

Tabela 3. Matrizes de confusão dos modelos gerados pelos algoritmos

KNN	Bayes	DTree	R.Forest	SVM	MLP
27 2 3 29	29 3 4 31	27 2 8 24	26 3 4 28	30 5 4 37	28 4 4 31

O Experimento

Tabela 2. Desempenho dos modelos gerados pelos algoritmos

	Acurácia	F1	TVP	TFP
KNN	91.8	92.06	93.10	9.38
Bayes	89.55	89.85	90.63	12.12
DTREE	83.6	82.75	93.10	25.00
Random Forest	88.52	88.88	89.66	12.50
SVM	88.15	89.15	85.7	9.76
MLP	88.06	88.57	87.50	12.12

Trabalhos Futuros

- Datasets com dados do Brasil(FIOCRUZ, HUAP)
- AHP
- SynthPop

FIM

Muito Obrigado!!!