Sistemas Inteligentes Aplicados

Trabalho: Introdução ao KDD e pré-processamento de dados

Parte 1 – Execute os treinamentos (peso 3)

1.1 Escolha 3 dentre os seguintes possíveis algoritmos de aprendizado de máquina:

- Árvore de Decisão CART
- Support Vector Machine
- k-NN
- Gaussian Naive Bayes
- Categorical Naive Bayes
- Regressão Linear (quadrados mínimos)
- Perceptron
- Árvore de decisão

1.2 Conforme o algoritmo, é possível que atributos qualitativos ou quantitativos não sejam aceitos. Vocês vão ter que definir se o algoritmo é adequado para os atributos que tem ou não. Se necessário, fazer as conversões/normalizações/discretizações de dados, ou ainda em caso extremos, a remoção do atributo (justificada pelas características do algoritmo e/ou métodos de seleção de atributos). Verifique também a possível necessidade de balanceamento de dados.

1.3 Treinar cada um dos algoritmos com os dados modificados para tal.

1.4 Escolher dos parâmetros pelo menos um parâmetro de cada algoritmo e modificá-lo com, pelo menos, dois valores diferentes (lembrando que vai ter que justificar essas escolhas). Isto dá um total de, no mínimo, 6 resultados de classificação (2 para cada algoritmo escolhido).

Parte 2 – Elabore o relatório (peso 4)

Faça um relatório com no mínimo 1000 palavras (usar “contar palavras” do Word ou similar) contendo:

2.1 Uma breve descrição do dataset.

2.2 Uma descrição sobre as transformações aplicadas nos dados (por algoritmo, se necessário), justificando o motivo da necessidade das transformações.

2.3 Os resultados obtidos para cada teste realizado.

2.4 O que fazem os parâmetros alterados para nos itens 1.3 e 1.4, e quais valores foram usados.

2.5 Definição do melhor resultado.

Tabelas, figuras, legendas de tabelas e figura, capa, contra-capa, fórmulas, pseudo-código, etc não contam como palavras.

Não há formato obrigatório, se estiver em dúvida sobre qual formato usar, use report (no latex) ou abnt (latex, word, etc)

Enviar .tex ou .docx ou equivalente para contagem de palavras, não apenas o .pdf.

Parte 3 – Apresentação (peso 3)

3.1 – Desenvolver uma apresentação com no máximo 15 slides abordando de maneira a apresentar para a turma os dados do relatório desenvolvido na parte 2.

Avaliação

1. O trabalho pode ser feito em equipes de até 3 pessoas.

2. Detecção de plágio, mesmo que em pequenas porções de texto ou slides **anularão o trabalho**. Se estiver em dúvida sobre o que caracteriza plágio, consulte o professor.

3. Entrega e apresentações **02/05/2024**

DataSet

As doenças cardiovasculares (DCV) são a causa número 1 de morte em todo o mundo, ceifando cerca de 17,9 milhões de vidas a cada ano, o que representa 31% de todas as mortes em todo o mundo. Quatro em cada cinco mortes por DCV são devidas a ataques cardíacos e acidentes vasculares cerebrais, e um terço destas mortes ocorre prematuramente em pessoas com menos de 70 anos de idade. A insuficiência cardíaca é um evento comum causado por DCV e este conjunto de dados contém 11 características que podem ser usadas para prever uma possível doença cardíaca.

Pessoas com doenças cardiovasculares ou que apresentam alto risco cardiovascular (devido à presença de um ou mais fatores de risco, como hipertensão, diabetes, hiperlipidemia ou doença já estabelecida) necessitam de detecção e gestão precoces, onde um modelo de aprendizagem automática pode ser de grande ajuda.

Informações de Atributo

Idade: idade do paciente [anos]

Sexo: sexo do paciente [M: Masculino, F: Feminino]

ChestPainType: tipo de dor torácica [TA: Angina típica, ATA: Angina atípica, NAP: Dor não anginosa, ASY: Assintomática]

PA em repouso: pressão arterial em repouso [mm Hg]

Colesterol: colesterol sérico [mm/dl]

BS em jejum: glicemia em jejum [1: se BS em jejum > 120 mg/dl, 0: caso contrário]

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]

MaxHR: frequência cardíaca máxima alcançada [valor numérico entre 60 e 202]

ExercícioAngina: angina induzida por exercício [Y: Sim, N: Não]

Oldpeak: oldpeak = ST [Valor numérico medido na depressão]

ST_Slope: a inclinação do segmento ST do pico do exercício [Up: subida, Flat: flat, Down: downsloping]

HeartDisease: classe de saída [1: doença cardíaca, 0: Normal]

Fonte: 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

Cada conjunto de dados usado pode ser encontrado no Índice de conjuntos de dados de doenças cardíacas do UCI Machine Learning Repository no seguinte link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Fontes de dados reais.

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.