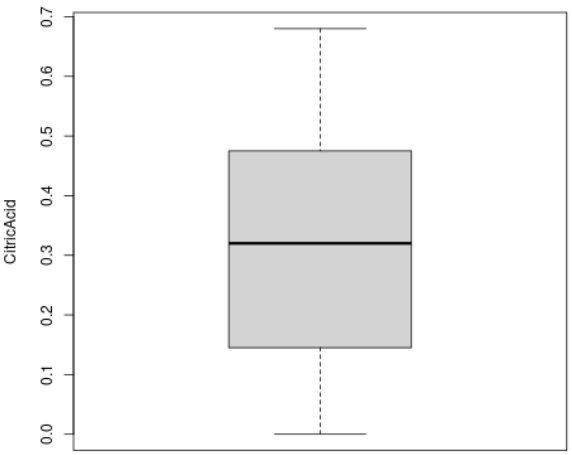
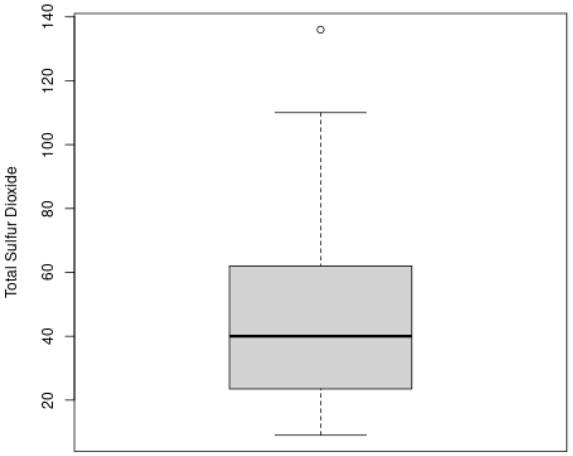
Atividade prática: Algoritmo k-Nearest Neighbors (KNN)

Objetivo da atividade:

- Compreender o processo de classificação através do algoritmo KNN
- Analisar o efeito de aspectos como normalização de dados e dimensão do vetor de atributos sobre a saída prevista pelo classificador

1. Faça o download dos dados disponibilizados no Moodle. Estes dados referem-se a uma tarefa de prever qualidade de vinhos tintos a partir da avaliação de características fisico-químicas analisadas. Os dados originais[1] foram pré-processados para que a tarefa seja de classificação binária, indicando vinhos de alta qualidade (class=1) ou de baixa qualidade (class=0), e para obter apenas uma pequena amostra dos dados (44 instâncias de treinamento, 4 instâncias de teste). Todos os atributos são numéricos, mas variam em escala. Por exemplo, para dióxido de enxofre total e ácido cítrico, temos as seguintes distribuições:





Cada instância possui um identificador único, disposto na coluna "ID".
Atenção: Este atributo **não** deve ser usado na construção do modelo KNN.

[1] https://archive.ics.uci.edu/ml/datasets/wine+quality

2. Treine um modelo utilizando o algoritmo KNN com a sua linguagem de programação ou software de preferência seguindo o guia de experimentos abaixo. A partir dos seus resultados para cada item, responda as questões no questionário sobre o KNN disponível no Moodle da disciplina.

- Você pode utilizar implementações prontas ou fazer a sua própria implementação, entretanto, é importante que seja capaz de recuperar as distâncias calculadas e índices/IDs dos K-vizinhos mais próximos pois os mesmos deverão ser reportados nos resultados.
- Utilize sempre a distância euclidiana como métrica de distância e a acurácia para avaliação dos modelos.
- Os dados de treinamento e teste são disponibilizados em arquivos distintos e devem ser usados conforme instruções a seguir. **Não devem** ser feitas outras divisões dos dados usando o método holdout, cross-validation, etc.

A estrutura das pastas fornecidas é explicada abaixo:

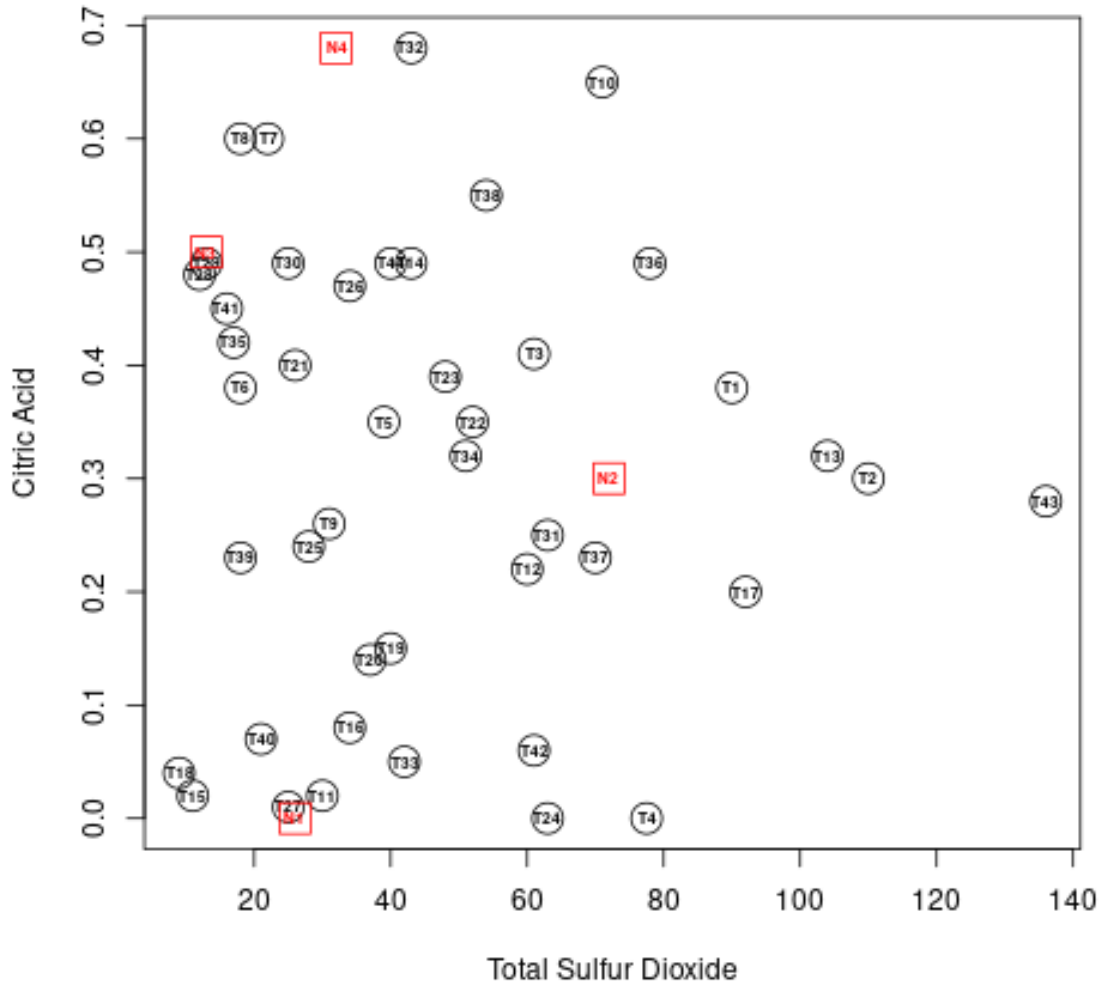
- `Dados_Originais_2Features`
  - Dados sem normalização, contendo apenas dois atributos selecionados, *dióxido de enxofre total* e *ácido cítrico*. Dados separados em TrainingSet (44 instâncias) e TestingSet (4 instâncias). O arquivo TestingSet contém a coluna 'class' com a classe correta de cada instância para viabilizar a avaliação do modelo por meio da acurácia.
- `Dados_Normalizados_2Features`
  - Dados com normalização pelo método min-max (ver slides da disciplina), seguindo a mesma estrutura do item anterior.
- `Dados_Originais_11Features`
  - Dados sem normalização, contendo os 11 fatributos disponibilizados originalmente. Dados separados em TrainingSet (44 instâncias) e TestingSet (4 instâncias). O arquivo TestingSet contém a coluna 'class' com a classe correta de cada instância para viabilizar a avaliação do modelo por meio da acurácia.
- `Dados_Normalizados_11Features`
  - Dados com normalização pelo método min-max (ver slides da disciplina), seguindo a mesma estrutura do item anterior.

Guia de experimentos

A. Treine modelos usando o algoritmo KNN para o conjunto de dados `Dados_Originais_2Features`, variando o valor de k (número de vizinhos mais próximos) entre 1, 3, 5, 7. Para cada modelo treinado, avalie seu desempenho nos dados de teste, reportando a acurácia. Repita o procedimento com os dados

`Dados_Normalizados_2Features` e compare as acurácias obtidas entre os dois modelos.

Visualização do conjunto de dados `Dados_Originais_2Features`. Os pontos representados como quadrados vermelhos referem-se às instâncias de teste.



B. Considerando o modelo treinado com k=5 utilizando dados não normalizados, verifique quem são os k vizinhos mais próximos da instância de teste **N1** (liste os respectivos IDs)**.** Verifique como estes vizinhos estão dispostos no espaço de entrada em relação à instância de teste N1 e aos eixos x e y. Após tirar suas conclusões, analise se as mesmas se aplicam às instâncias de teste N2, N3 e N4.

C. Treine dois modelos usando o algoritmo KNN com k=5 para os datasets `Dados_Normalizados_2Features` e `Dados_Normalizados_11Features`. Aplique os modelos treinados nos dados de teste, verificando os k-vizinhos mais próximos e a classe prevista para a instância **N4**. Faça perturbações no valor do atributo “citric acid” para a instância N4, substituindo o valor original (1.0) por **0.3** e posteriormente por **0.85** (ou seja, gere duas novas instâncias sintéticas com esta alteração). Repita a classificação destas instâncias sintéticas com os dois

modelos (isto é, modelo baseado em 2 atributos e en 11 atributos) . Compare os resultados, analisando como a alteração de um atributo impactou o cálculo das distâncias euclidianas e a seleção dos k-vizinhos mais próximos em cada caso.

O prazo final de entrega deste exercício é dia **30 de janeiro às 23:59h**. As respostas serão submetidas via questionário no Moodle.