Relatório - Análise de base de dados

Objetivo

Realizar a análise de dados a partir de uma base no formato *.csv* para a disciplina de **Desenvolvimento e execução de projetos de software** ministrada pelo professor **Ryan Azevedo** no semestre 2022.2**.**

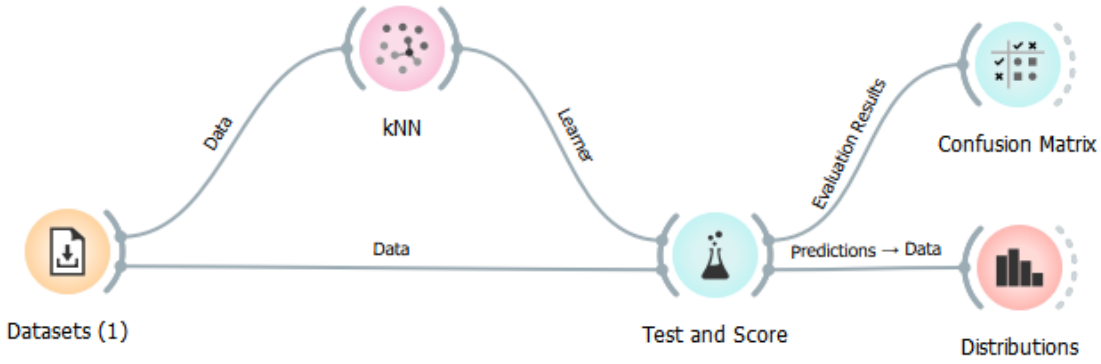
Integrantes

- Erik Alexandre
- Jackson Lima
- Luiz Felipe
- Victor Mendes

Ferramentas utilizadas

Orange Data Mining - Software open source para aprendizado de maquina e análise de dados.

Diagrama

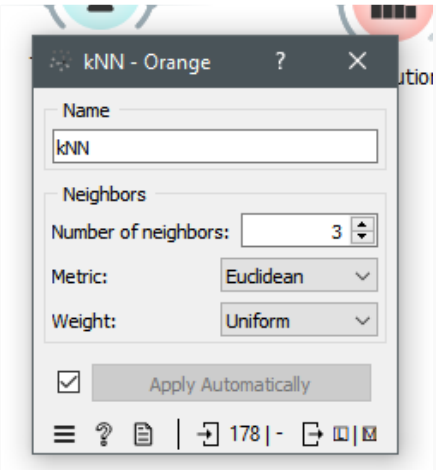


Análise com Orange KNN

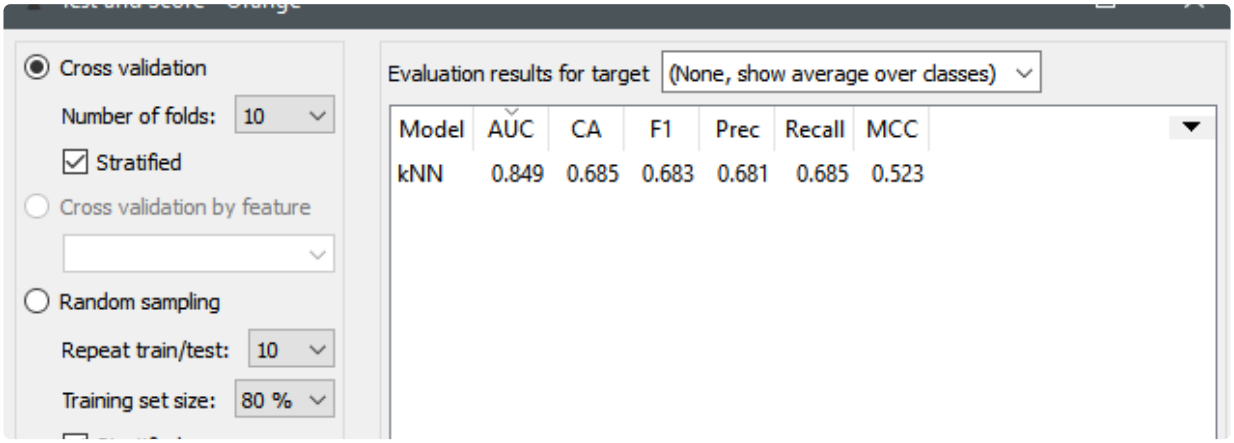
Foi utilizado uma base de dados chamada *load_wine* do *scikit-learn,* que importa um conjunto de dados de vinho para classificação deste dados.
De algoritmo foi utilizado o KNN, algoritmo de aprendizado de maquina supervisionado utilizado para classificação e regressão. É um dos algoritmos mais simples e amplamente utilizados no campo de aprendizado de máquina.

Para análise, passamos o seguinte parâmetros:

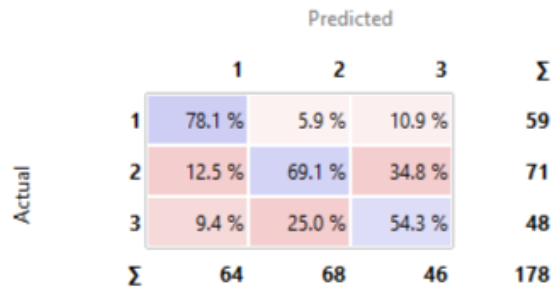
- Neighbors = 3
- Métrica = Euclidiana
- Altura = Uniforme



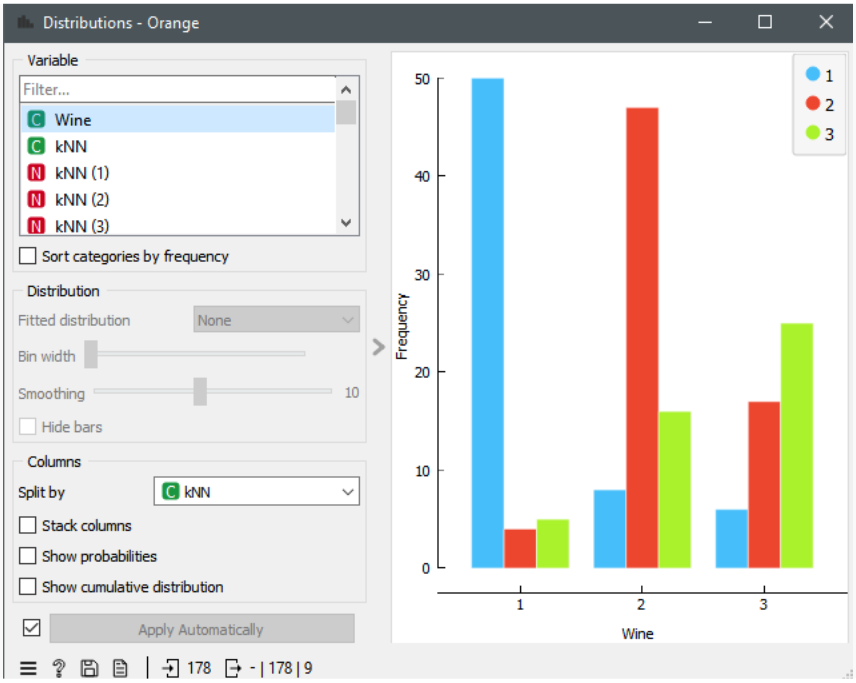
Com base nesses parâmetros, foi possível obter o seguinte resultado utilizando o *Test and Score*:



- Acurácia = 0.849
- Precissão = 0.681
- Recall = 0.685



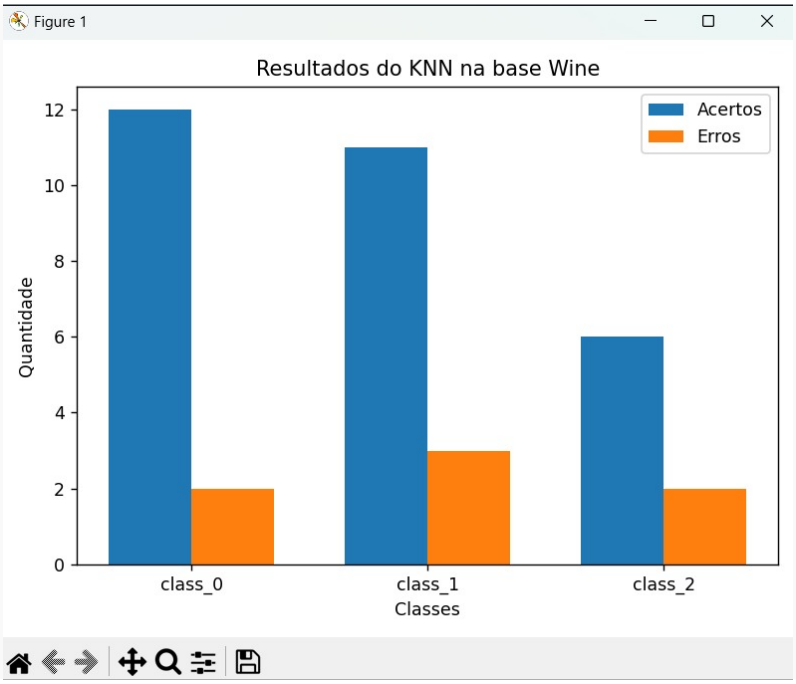
E, utilizando o *Distributions* conseguimos plotar um gráfico com base nesses dados:



Com isso, conseguimos representar a classificação de wine em cada vizinho do knn, que no caso são 3.

Análise com python

Ao realizar o KNN em python, foi possível obter o seguinte gráfico:



- Acurácia = 0.805
- Precisão = 0.823
- Recall = 0.805

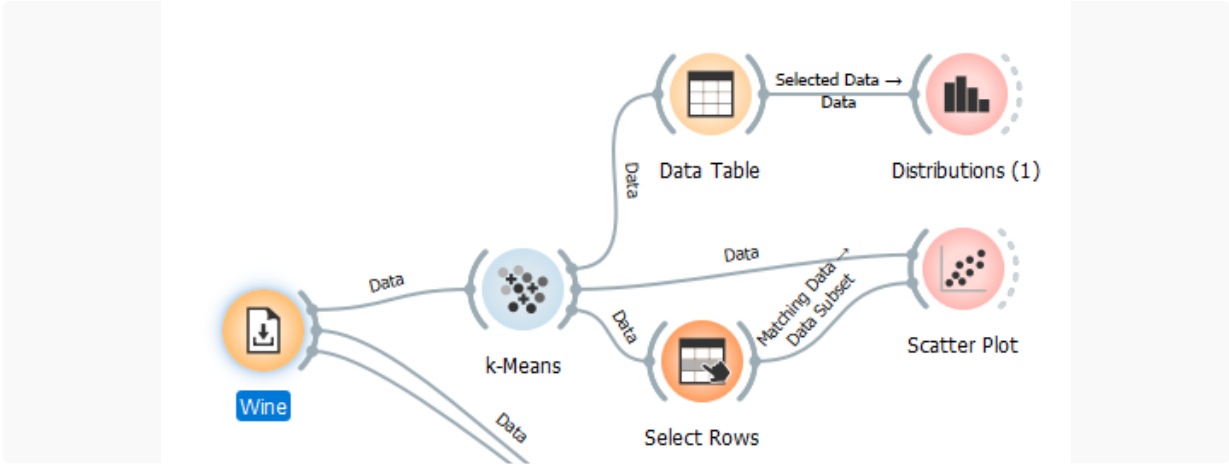
Comparação

	Acurácia	Precisão	Recall
Orange	0.849	0.681	0.685
Python	0.805	0.823	0.805
Diff	0.44	0.142	0.120

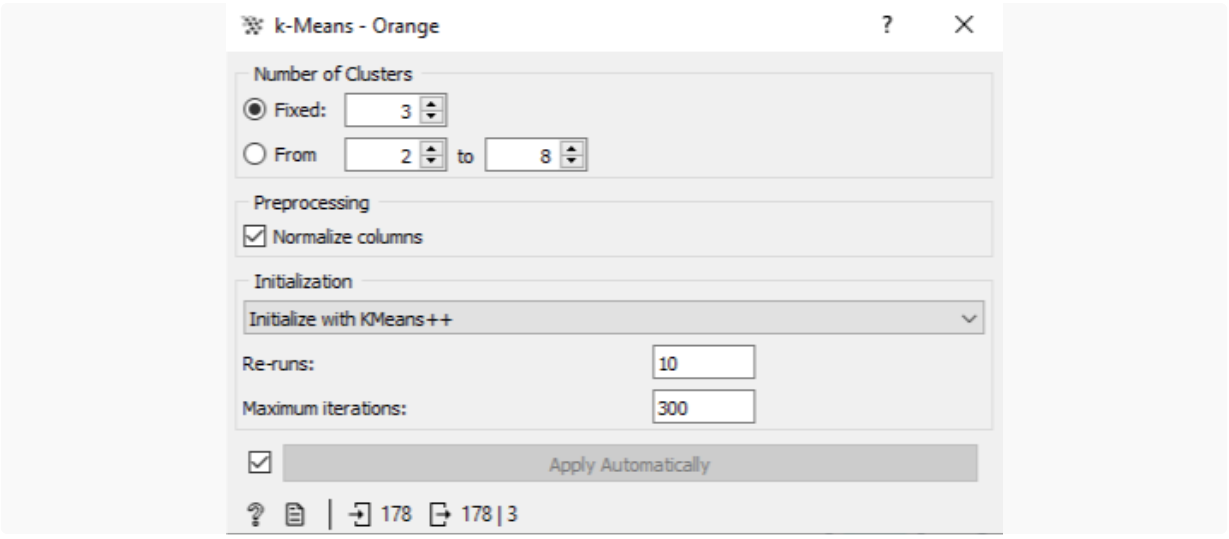
Análise no Orange com K-Means

Também foi realizado um teste utilizando a base wine, com o algoritmo k-Means, com o numero de clusters setado em 3, já que são utilizados 3 classes de wine.
O K-Means é um algoritmo de aprendizado não supervisionado usado para agrupar dados em clusters. Ele busca dividir um conjunto de dados em grupos distintos, onde os objetos dentro de um mesmo grupo são mais semelhantes entre si do que com os objetos em outros grupos.

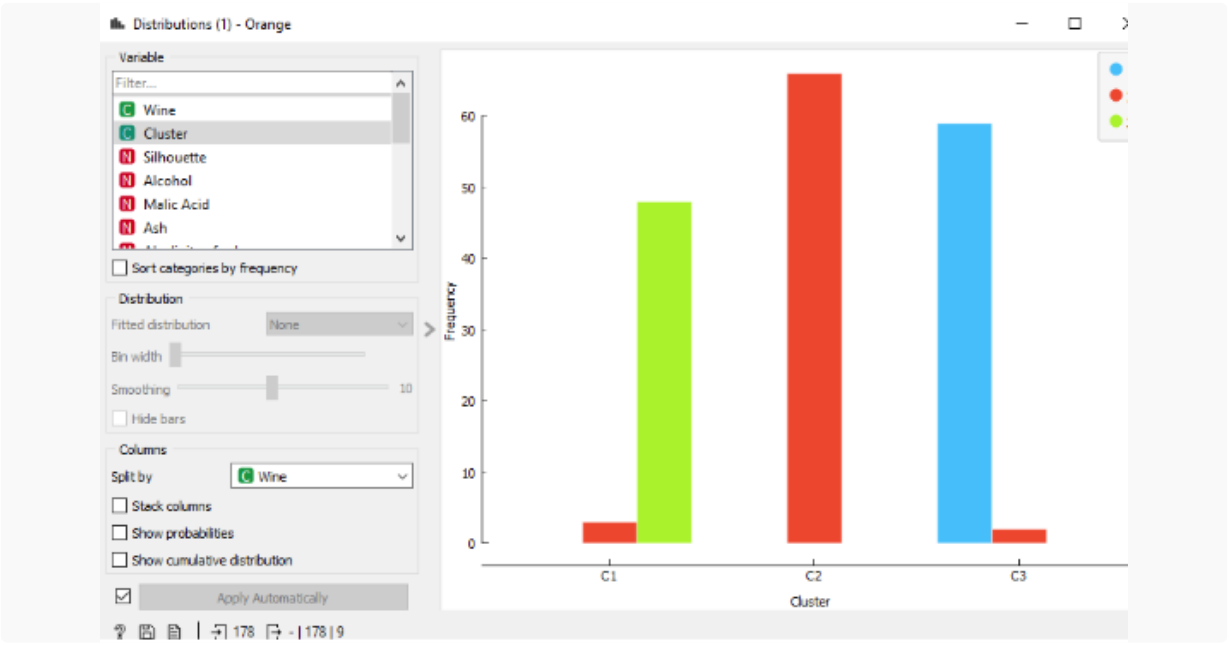
Diagrama



Dados

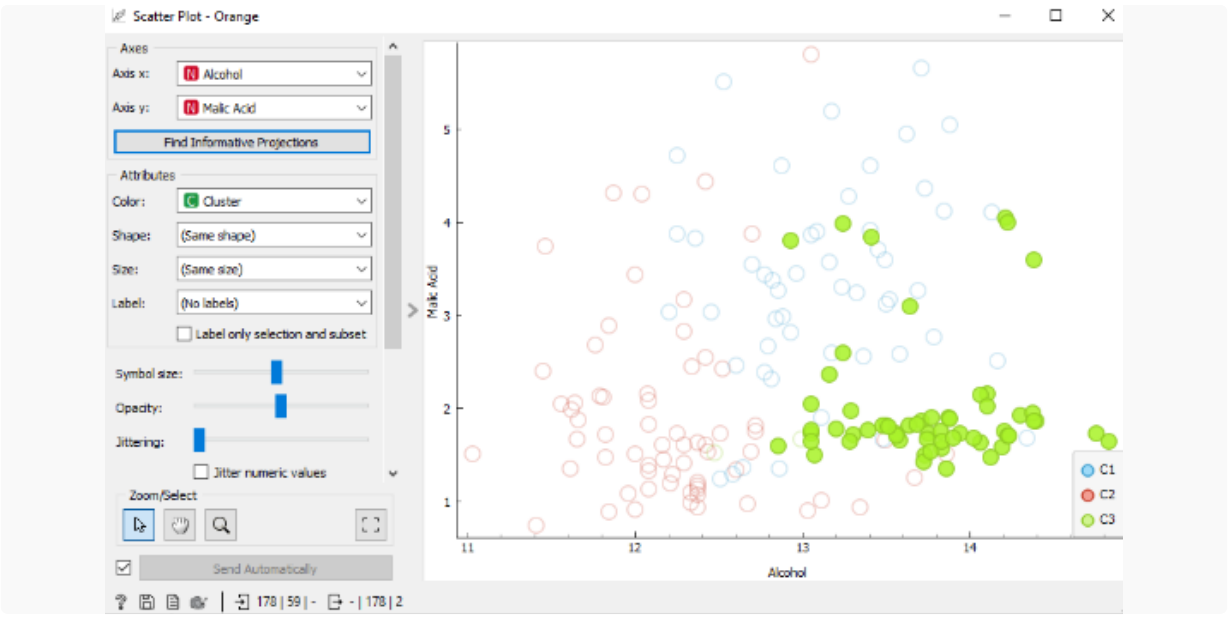


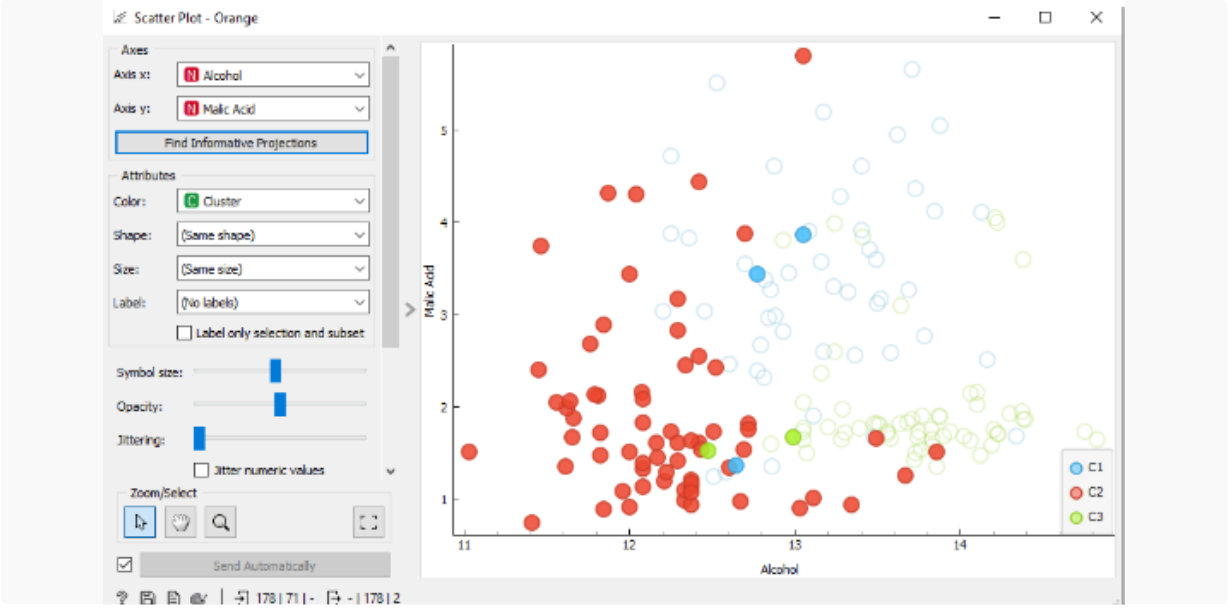
Resultado obtido com Distributions

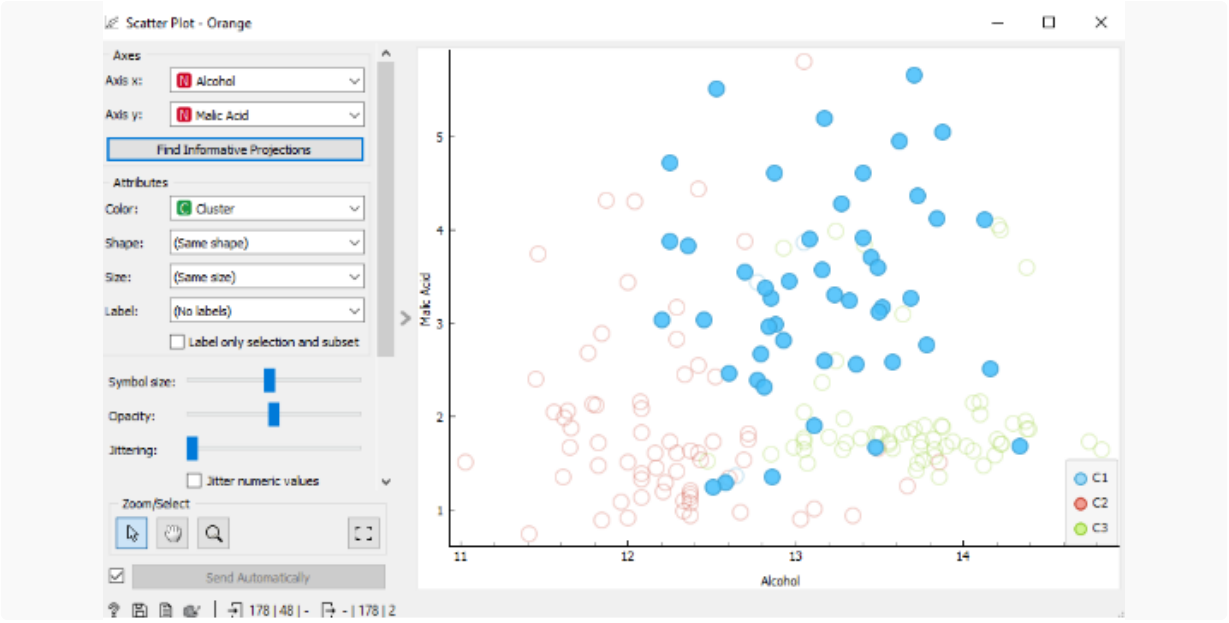


É possível analisar no gráfico que tanto o wine 1 quanto o wine 3 tiveram sua classificação correta porém, o wine 2 teve alguns elementos classificados no cluster 1 e no cluster 3.

É possível visualizar melhor esses dados nos gráficos ***Scatter*** a seguir


Scatter Plot - Orange
Axes
Axis x: Alcohol
Axis y: Malic Acid
Find Informative Projections
Attributes
Color: Cluster
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Zoom/Select
Send Automatically
178 | 59 | - - | 178 | 2
Malic Acid
Alcohol
C1
C2
C3


Scatter Plot - Orange
Axes
Axis x: Alcohol
Axis y: Malic Acid
Find Informative Projections
Attributes
Color: Cluster
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Zoom/Select
Send Automatically
178 | 71 | - - | 178 | 2
Malic Acid
Alcohol
C1
C2
C3



Como obtivemos 5 erros, então temos que $5/178 = 0,028089887640449$

Logo a taxa de acerto do knn foi de cerca de $0,98$