Machine Learning (P02)

Artificial Intelligence, 2022-23

**João Apresentação (21152), Pedro Simões (21140), Gonçalo Cunha (21145)**

**Conteúdo**

[1. Introdução 3](#_Toc124798692)

[1.1. Contexto 3](#_Toc124798693)

[1.2. Objetivos 3](#_Toc124798694)

[1.3. Estrutura do documento 3](#_Toc124798695)

[1.4. Data set (Iris Species) 3](#_Toc124798696)

[1.4.1. Descrição 3](#_Toc124798697)

[1.4.2. Meta data 3](#_Toc124798698)

[1.4.2.1. Colaboradores 3](#_Toc124798699)

[1.4.2.2. Licença 3](#_Toc124798700)

[1.4.2.3. Frequência de atualização esperado 3](#_Toc124798701)

[2. Classificação automática 4](#_Toc124798702)

[2.1. Objetivos de negócio a alcançar 4](#_Toc124798703)

[2.2. Algoritmos e parâmetros selecionados 4](#_Toc124798704)

[2.2.1. SVM 4](#_Toc124798705)

[2.2.2. Naive Bayes 5](#_Toc124798706)

[2.2.3. Random Forest 5](#_Toc124798707)

[2.3. Critérios de seleção de dados e preparação dos dados 6](#_Toc124798708)

[2.4. Avaliação dos modelos de classificação 7](#_Toc124798709)

[2.5. Resultados 8](#_Toc124798710)

[3. Clustering 9](#_Toc124798711)

[3.1. Objetivos de negócio a alcançar 9](#_Toc124798712)

[3.2. Critérios de seleção de dados e preparação dos dados 9](#_Toc124798713)

[3.3. Avaliação da aplicação do algoritmo K-Means 10](#_Toc124798714)

[4. Regras de Associação 11](#_Toc124798715)

[4.1. Objetivos de negócio a alcançar 11](#_Toc124798716)

[4.2. Critérios de seleção de dados e preparação dos dados 11](#_Toc124798717)

[4.3. Resultados da avaliação da aplicação do algoritmo Apriori 12](#_Toc124798718)

[5. Conclusão 13](#_Toc124798719)

[6. Bibliografia 13](#_Toc124798720)

# Introdução

## Contexto

Este trabalho prático, relativo à unidade curricular de **Inteligência Artificial,** propende desenvolver um programa que lê uma Dataset e aplica algoritmos de Machine Learning para classificação, clustering e regras de associação.

Para o desenvolvimento foi utilizado o Knime para a classificação e clustering e Orange para a regras de associação.

## Objetivos

* Implementar e analisar diferentes abordagens de Machine Learning;
* Métodos para resolver um problema específico usando um conjunto de dados aberto/público.

## Estrutura do documento

O documento está estruturado de forma que seja de simples leitura. Existe recurso a referências de material fornecido pelo professor Joaquim Silva e/ou referências a excertos de Web grafia.

## Data set (Iris Species)

Foi escolhida esta Dataset tendo em conta a sua fácil interpretação e dados adequados ao trabalho proposto.

Este foi utilizado em aula.

## Descrição

O Data set selecionado para este projeto é de uma determinação da espécie da flor Iris, dividida em 3 espécies (Iris-septosa, Iris-versicolor, Iris-virginica).

A data set apresenta reduzidas caraterísticas, dividindo-se em 6 colunas:

* Id;
* SepalLenghtCm (comprimento da sépala em cm);
* SepalWidthCm (largura da sépala em cm);
* PetalLengthCm (comprimento da pétala em cm);
* PetalWidthCm (largura da pétala em cm);
* Species (espécie) -> Categórico.

## Meta data

## Colaboradores

UCI Machine Learning (Owner)

## Licença

[CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/)

## Frequência de atualização esperado

Não especificado (Atualizado 6 anos atrás)

# Classificação automática

Para a realização deste tópico foi utilizada a ferramenta Knime.

## Objetivos de negócio a alcançar

**Seleção de espécies:** O modelo de classificação automática pode ser usado para selecionar espécies de Iris com características desejadas para reprodução ou venda. Isso pode ajudar a melhorar a qualidade e a rentabilidade da produção.

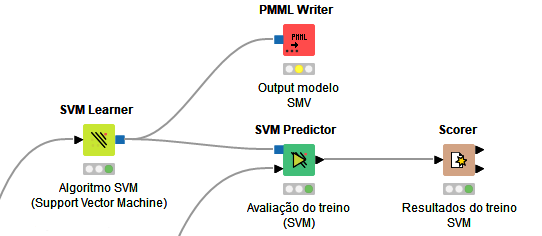
## Algoritmos e parâmetros selecionados

Uma razão para testar a classificação automática com Naive Bayes, Random Forest e SVM é para avaliar qual algoritmo tem o melhor desempenho no Dataset de espécies de Iris. Cada algoritmo tem suas próprias vantagens e desvantagens e pode se sair melhor ou pior em diferentes conjuntos de dados e situações.

Comparando esses algoritmos, é possível avaliar qual é o melhor para classificar as espécies de Iris no Dataset específico. Isso pode ser útil para determinar qual algoritmo usar para um determinado problema.

## SVM

SVM é conhecido por sua eficácia na classificação e capacidade de lidar com dados de alta dimensionalidade, mas pode ser mais difícil de interpretar.

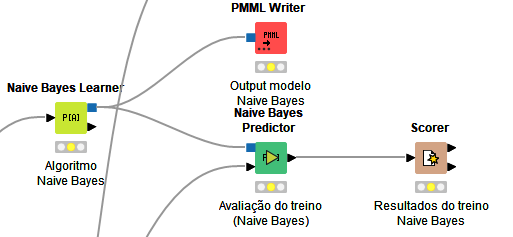


Knime - Aprendizagem SVM

## Naive Bayes

Naive Bayes é conhecido por ser rápido e eficiente com poucos dados, mas pode ser menos preciso quando há muitas características.

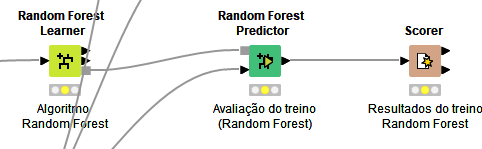
O que é favorável para esta Dataset tendo em conta que ela possui apenas 4 caraterísticas (atributos).



Knime 1 - Aprendizagem Naive Bayes

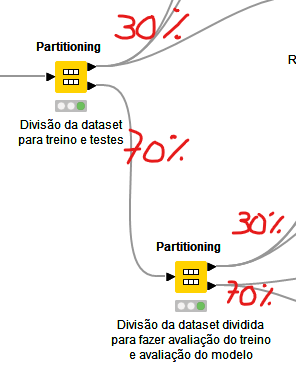
## Random Forest

Random Forest é um algoritmo robusto e preciso, mas pode ser mais lento e requerer mais dados.



Knime 2 - Aprendizagem Random Forest

## Critérios de seleção de dados e preparação dos dados



Knime 3 - Partição do Dataset

Inicialmente foi discretizado a coluna referente ao id, tendo em conta o seu reduzido impacto na classificação

.

Foi selecionado da data set inicial, **30%** para **treino** e **70%** para **testes**.

A razão para ter uma maior percentagem de dados no conjunto de teste do que no conjunto de treino é para fornecer uma amostra suficientemente grande de dados para avaliar a capacidade de generalização do modelo. Quanto maior o conjunto de teste, maior a precisão da avaliação do modelo. Além disso, um conjunto de teste maior também permite avaliar o desempenho do modelo em diferentes subconjuntos de dados, o que é útil para identificar tendências e problemas.

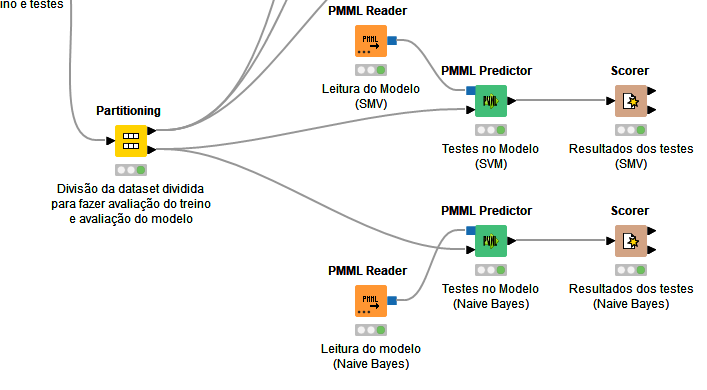
Desses 70% para testes, 30% são usados para fazer a avaliação da precisão dos treinos, e 70% para fazer a avaliação da precisão dos testes.

## Avaliação dos modelos de classificação

Encontram-se aqui os dois modelos de classificação, um do Naive Bayes e outro do SMV.

Estes são importados tendo em conta que após o treino de cada algoritmo é feita uma exportação em PMML.

Não foi feito um modelo de classificação para o Random Forest tendo em conta que a previsão final é baseada na média das previsões de cada árvore presente.



Knime 4 - PMML Predictor

## Resultados

O SVM terá obtido os piores resultados devido ao conjunto de dados possuir poucas caraterísticas.

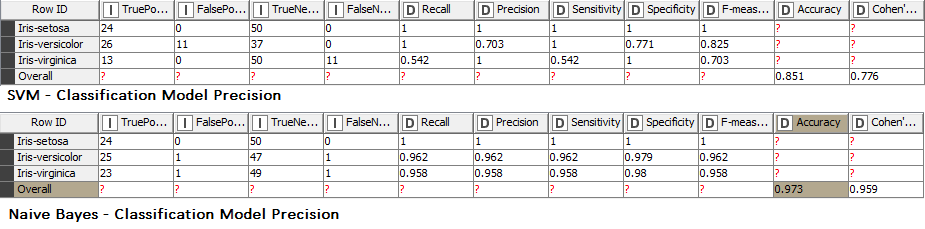
O Random Forest terá obtido resultados melhores na fase de treinos, mas em resultados de modelo o Naive Bayes superou.

Na fase de avaliação dos modelos de classificação, o Naive Bayes teve o melhor resultado para previsão dos dados, com uma precisão de 97.3%, devido á Dataset apresentar um elevado numero de dados e reduzido numero de caraterísticas.

Uma imagem com texto, armário, captura de ecrã

Descrição gerada automaticamente

Knime 5 – Resultados dos algoritmos de treino



Knime 6 - Resultados dos modelos de classificação

Foi possível também obter as matrizes de confusão de cada avaliação de precisão (tanto nos testes como para modelos), segue-se abaixo um exemplo de uma matriz de confusão na avaliação de precisão do modelo de classificação de Naive Bayes:

Uma imagem com mesa

Descrição gerada automaticamente

Knime 7 - Matriz de confusão Naive Bayes Predictor

Neste exemplo verifica-se a existência de um falso positivo e falso negativo, no caso da previsão de uma Iris-virginica e obteve-se Iris-versicolor e o resto dos resultados terem sido verdadeiros positivos.

# Clustering

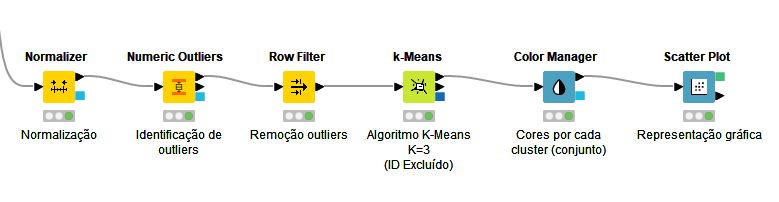
Para a realização deste tópico foi utilizada a ferramenta Knime.

## Objetivos de negócio a alcançar

**Previsão de espécies:** O modelo de classificação automática treinado com o dataset de espécies de Iris pode ser usado para prever a espécie de uma planta de Iris com base em suas características, como comprimento e largura da sépala e pétala. Isso pode ser útil para ajudar os jardineiros e vendedores a identificar corretamente as plantas de Iris.

## Critérios de seleção de dados e preparação dos dados

Para a montagem do clustering foi inicialmente normalizado os dados para todas as caraterísticas (exceto id) apresentarem uma mesma escala de valores entre 0 e 10. Foi identificado e removido linhas que incluíam outliers pois k-Means é sensível.



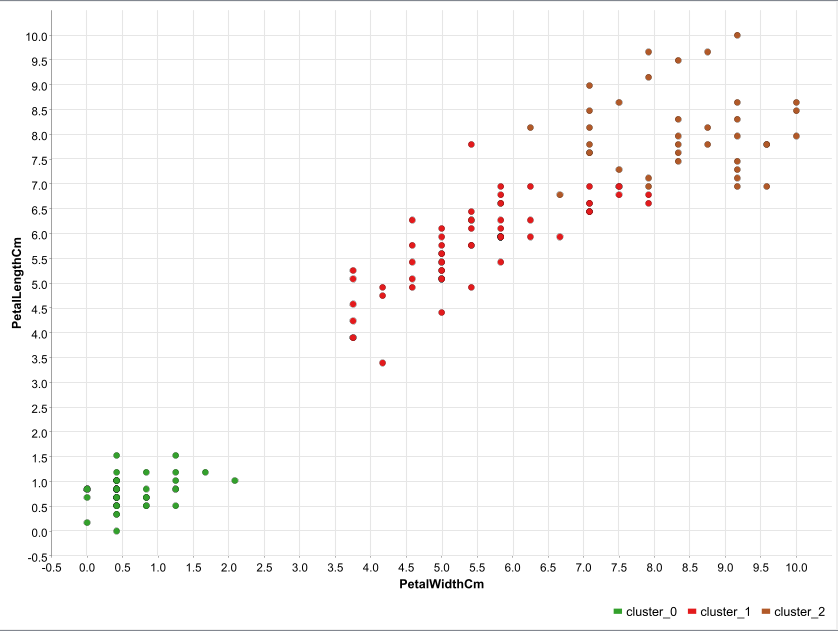
Knime 8 – Clustering

## Avaliação da aplicação do algoritmo K-Means

O k-Means recebe como parâmetros as caraterísticas todas (exceto o id) e o k.

O valor de k selecionado foi 3, tendo em conta que o objetivo deste clustering era fazer a divisão em 3 grupos de espécies de Iris presentes no atributo categórico.

Foi selecionada como número de iterações máximo 100, e este valor foi determinado após vários testes.



Knime 9 - Resultados Clustering

# Regras de Associação

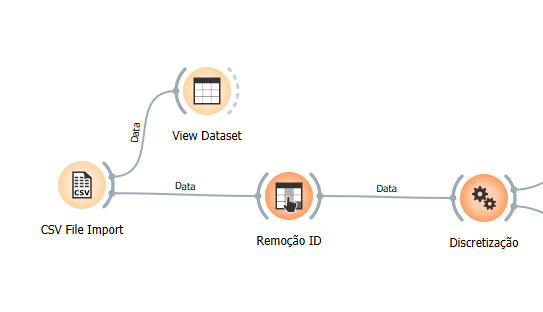
Para a realização deste tópico foi utilizada a ferramenta Orange Data Mining.

## Objetivos de negócio a alcançar

1. **Identificação de características importantes:** As regras de associação podem ser usadas para identificar quais características das espécies de Iris são mais importantes para a classificação das espécies. Isso pode ajudar a identificar quais características são as mais relevantes para distinguir as espécies.
2. **Identificação de relações entre características:** As regras de associação podem ser usadas para identificar relações entre as características das espécies de Iris, como quais características são frequentemente encontradas juntas em uma mesma espécie. Isso pode ajudar a entender como as características das espécies estão relacionadas entre si.

## Critérios de seleção de dados e preparação dos dados

É feita uma **remoção** do id e **discretização** de dados para estes serem utilizados na identificação de conjuntos de dados frequentes.



Orange 1 - Preparação dos dados

Os dados tiveram que ser discretizados tendo em conta que a Dataset possui variáveis continuas, mas para determinar o conjunto de dados frequentes é necessário trabalhar com variáveis discretas.

Esta operação resolve este problema.

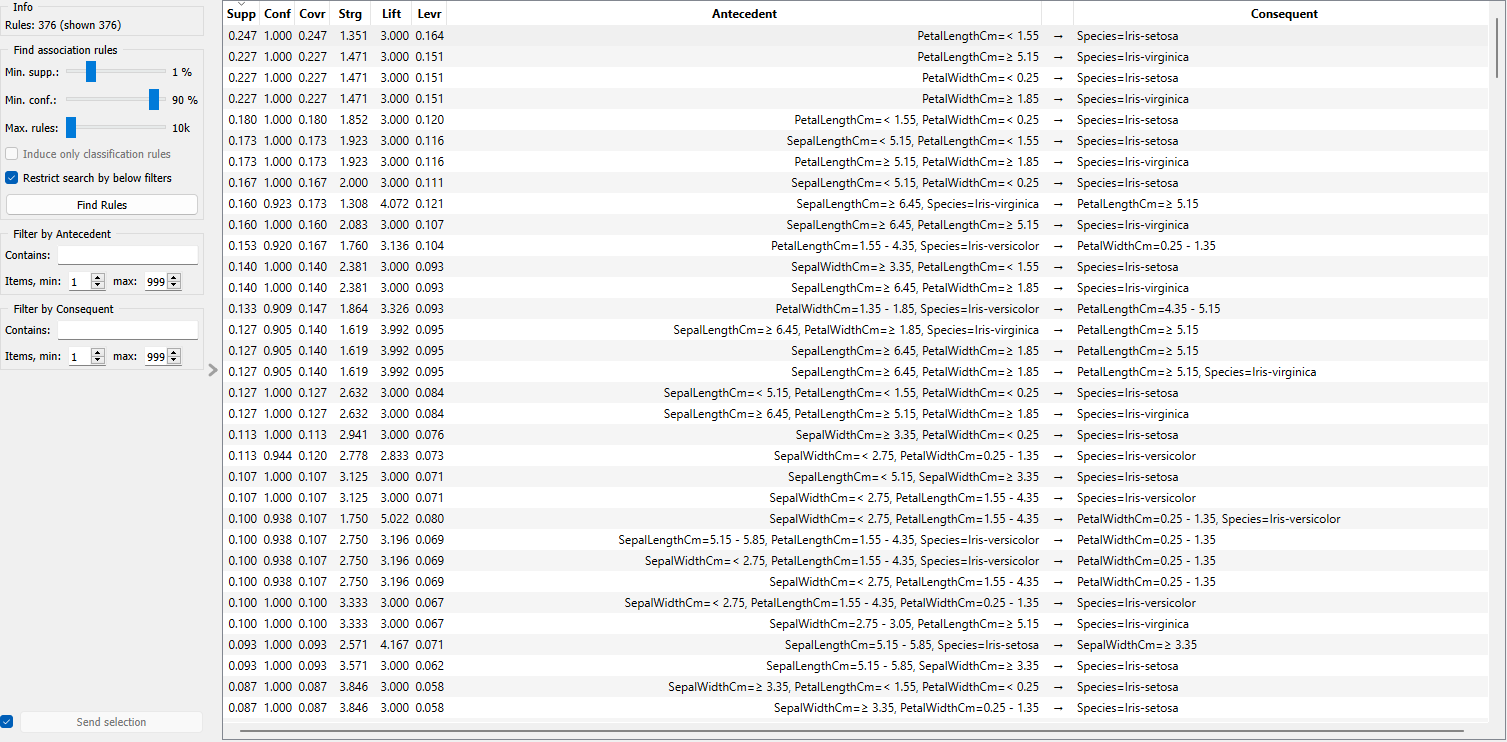
Uma imagem com texto

Descrição gerada automaticamente

Orange 2 – Discretização

## Resultados da avaliação da aplicação do algoritmo Apriori

Através da seguinte imagem é possível observar os resultados obtidos pelo algoritmo.



Orange 3 - Resultados Apriori

Nestas linhas de exemplo que se seguem é possível observar que:

1. Cerca de **24.7%** das Iris com comprimento de pétala menor a 1.55 cm, tem **100%** de grau de confiança para ser da espécie setosa;
2. Cerca de **22.7%** das Iris com comprimento de pétala maior ou igual a 5.15 cm, tem **100%** de grau de confiança para ser da espécie virginica;
3. Cerca de **22.7%** das Iris com largura de pétala menor que 0.25 cm, tem **100%** de grau de confiança para ser da espécie setosa.



Orange 4 - Resultados Apriori (exemplos)

# Conclusão

Com a elaboração deste pequeno projeto foi possível aplicar todas as aulas teóricas relacionadas com Machine Learning, desde algoritmos de aprendizagem à construção de modelos de classificação, clustering e regras de associação.

# Bibliografia

Iris Species Data set: <https://www.kaggle.com/datasets/uciml/iris>

Knime install: https://www.knime.com/downloads

Orange install: <https://orangedatamining.com>

Repositório GitHub: https://github.com/L0ud3r/MachineLearning