Machine Learning (P02)

Artificial Intelligence, 2022-23

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

Conteúdo

1. Introdução 3

1.1. Contexto 3

1.2. Objetivos 3

1.3. Estrutura do documento 3

1.4. Data set (Iris Species) 3

1.4.1. Descrição 3

1.4.2. Meta data 3

1.4.2.1. Colaboradores 3

1.4.2.2. Licença 3

1.4.2.3. Frequência de atualização esperado 3

2. Classificação automática 4

2.1. Objetivos de negócio a alcançar 4

2.2. Algoritmos e parâmetros selecionados 4

2.2.1. SVM 4

2.2.2. Naive Bayes 5

2.2.3. Random Forest 5

2.3. Critérios de seleção de dados e preparação dos dados 6

2.4. Avaliação dos modelos de classificação 6

2.5. Resultados 7

3. Clustering 8

3.1. Objetivos de negócio a alcançar 8

3.2. Critérios de seleção de dados e preparação dos dados 8

3.3. Avaliação da aplicação do algoritmo K-Means 9

4. Regras de Associação 10

4.1. Objetivos de negócio a alcançar 10

4.2. Critérios de seleção de dados e preparação dos dados 10

4.3. Resultados da avaliação da aplicação do algoritmo Apriori 11

5. Conclusão 12

6. Bibliografia 12

1. Introdução

1.1. 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.

1.2. Objetivos

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

1.3. 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.

1.4. 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.

1.4.1.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.

1.4.2.Meta data

1.4.2.1. Colaboradores

UCI Machine Learning (Owner)

1.4.2.2. Licença

CC0: Public Domain

1.4.2.3. Frequência de atualização esperado

Não especificado (Atualizado 6 anos atrás)

2. Classificação automática

Para a realização deste tópico foi utilizada a ferramenta Knime.

2.1. 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.

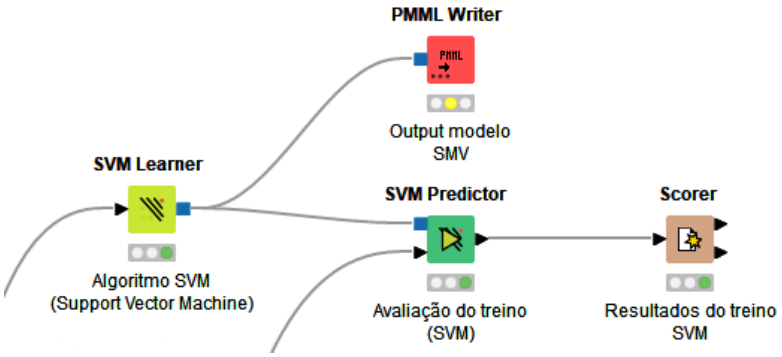
2.2. Algoritmos e parâmetros selecionados

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.

2.2.1.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.

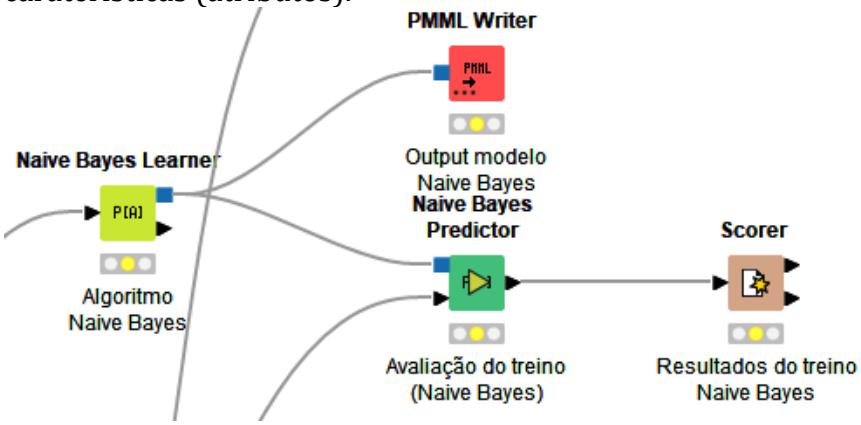


Knime - Aprendizagem SVM

2.2.2.Naive Bayes

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

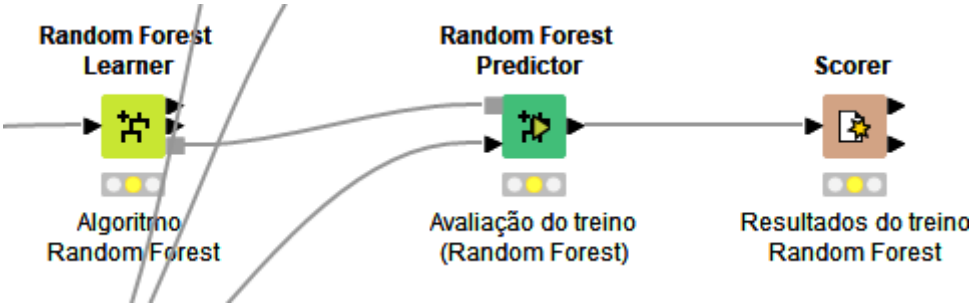
O que é favorável para esta Dataset tendo em conta que ela possui apenas 4 caraterísticas (atributos).



Knime 1 - Aprendizagem Naive Bayes

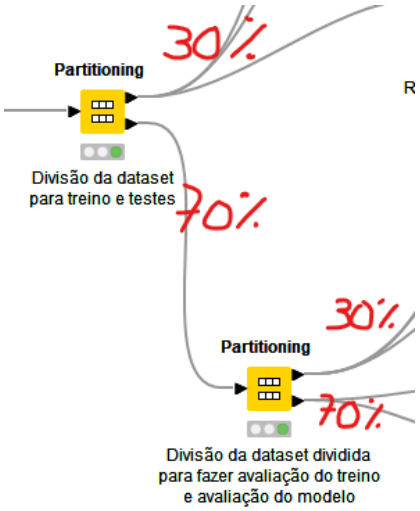
2.2.3.Random Forest

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



Knime 2 - Aprendizagem Random Forest

2.3. 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

.

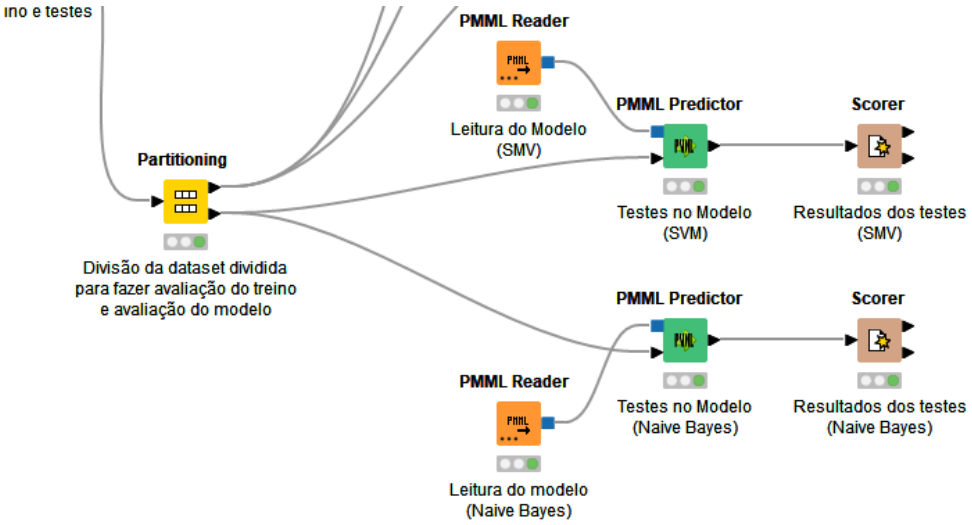
Foi selecionado da data set inicial, **30%** para **treino** e **70%** para **testes**.
A razão para ter uma maior percentagem de dados no conjunto de teste do que no conjunto de treino é para fornecer uma amostra suficientemente grande de dados para avaliar a capacidade de generalização do modelo. Quanto 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.

2.4. 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

2.5. 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.

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Iris-setosa	12	0	19	0	1	1	1	1	1	?	?
Iris-versicolor	7	1	20	3	0.7	0.875	0.7	0.952	0.778	?	?
Iris-virginica	8	3	19	1	0.889	0.727	0.889	0.864	0.8	?	?
Overall	?	?	?	?	?	?	?	?	?	0.871	0.806

SMV

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Iris-setosa	12	0	19	0	1	1	1	1	1	?	?
Iris-versicolor	9	2	19	1	0.9	0.818	0.9	0.905	0.857	?	?
Iris-virginica	7	1	21	2	0.778	0.875	0.778	0.955	0.824	?	?
Overall	?	?	?	?	?	?	?	?	?	0.903	0.854

Naive Bayes

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Iris-setosa	12	0	19	0	1	1	1	1	1	?	?
Iris-versicolor	9	1	20	1	0.9	0.9	0.9	0.952	0.9	?	?
Iris-virginica	8	1	21	1	0.889	0.889	0.889	0.955	0.889	?	?
Overall	?	?	?	?	?	?	?	?	?	0.935	0.903

Random-forest

Knime 5 – Resultados dos algoritmos de treino

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Iris-setosa	24	0	50	0	1	1	1	1	1	?	?
Iris-versicolor	26	11	37	0	1	0.703	1	0.771	0.825	?	?
Iris-virginica	13	0	50	11	0.542	1	0.542	1	0.703	?	?
Overall	?	?	?	?	?	?	?	?	?	0.851	0.776

SVM - Classification Model Precision

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Iris-setosa	24	0	50	0	1	1	1	1	1	?	?
Iris-versicolor	25	1	47	1	0.962	0.962	0.962	0.979	0.962	?	?
Iris-virginica	23	1	49	1	0.958	0.958	0.958	0.98	0.958	?	?
Overall	?	?	?	?	?	?	?	?	?	0.973	0.959

Naive Bayes - Classification Model Precision

Knime 6 - Resultados dos modelos de classificação

3. Clustering

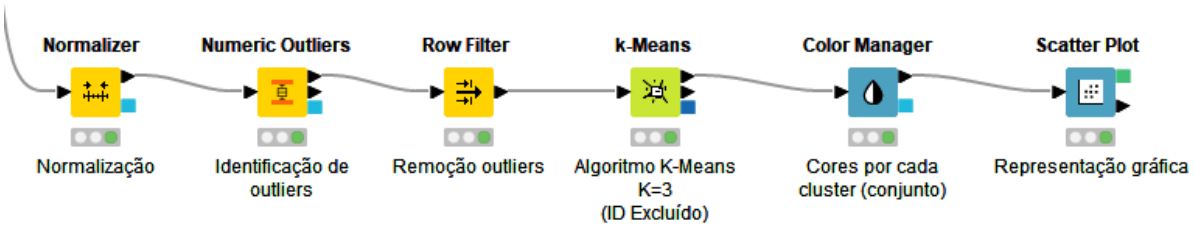
Para a realização deste tópico foi utilizada a ferramenta Knime.

3.1. 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.

3.2. 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.



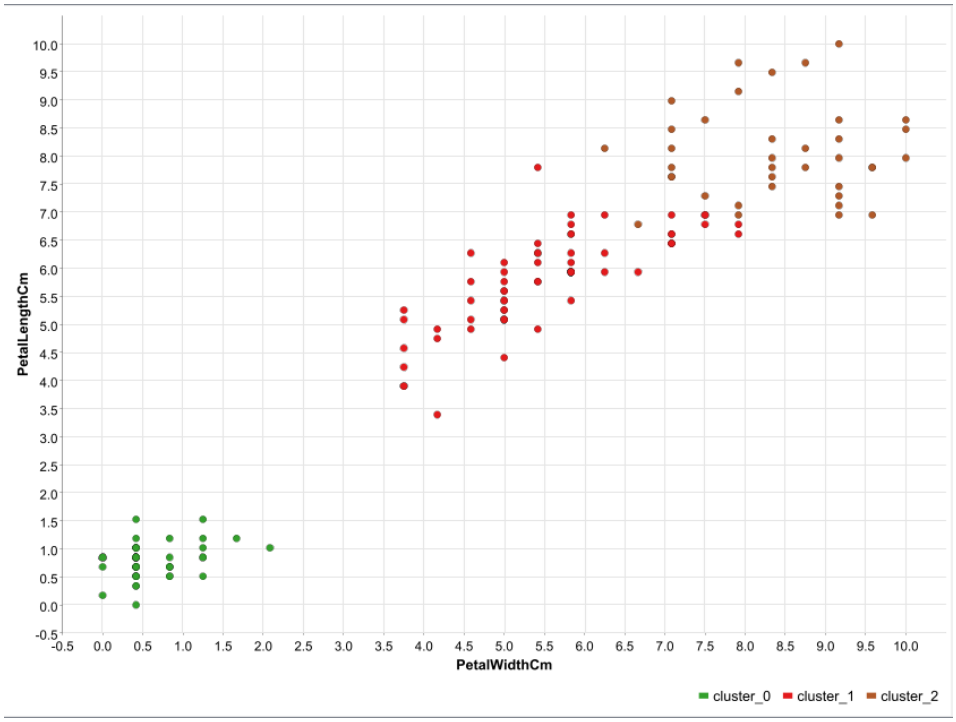
Knime 7 – Clustering

3.3. 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 8 - Resultados Clustering

4. Regras de Associação

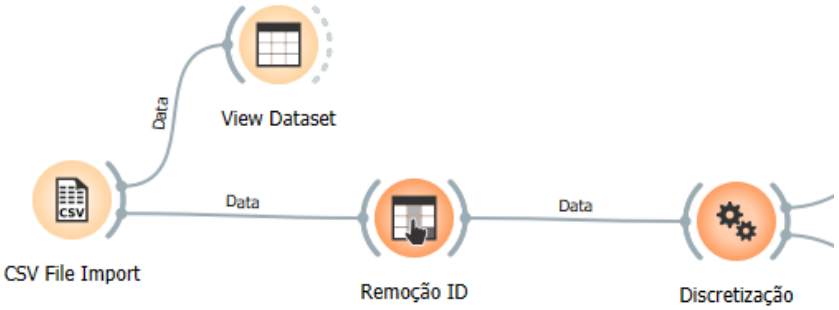
Para a realização deste tópico foi utilizada a ferramenta Orange Data Mining.

4.1. 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.

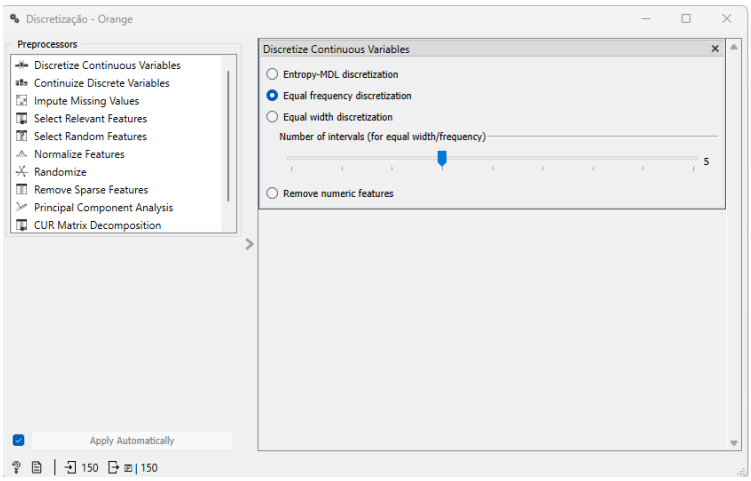
4.2. Critérios de seleção de dados e preparação dos dados

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

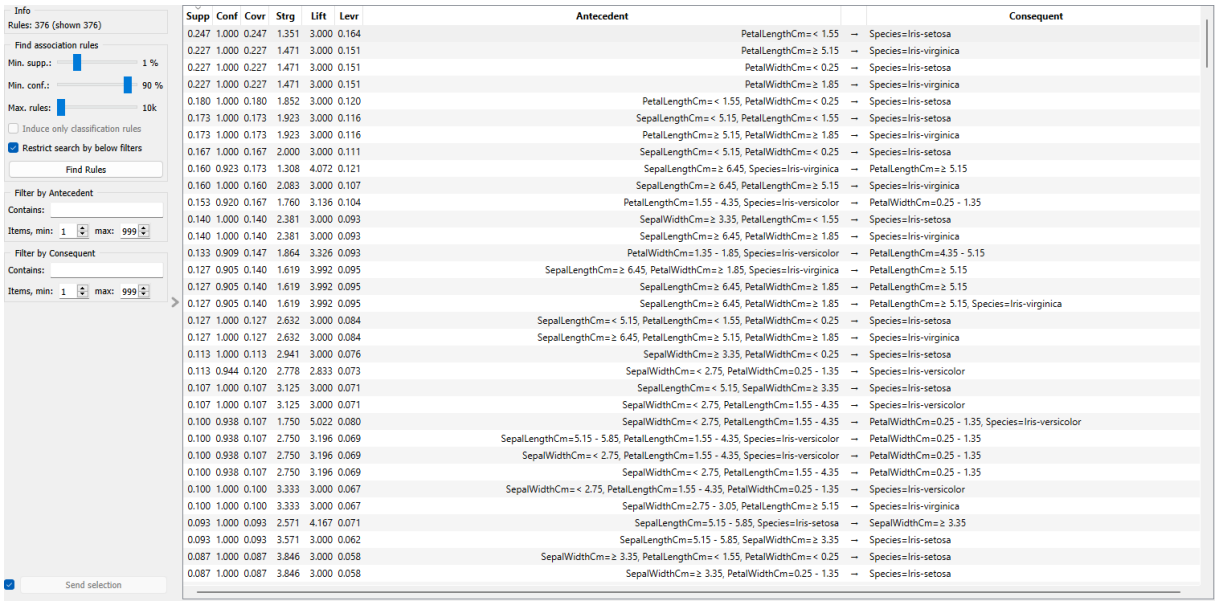
Os dados tiveram que ser discretizados tendo em conta que a Dataset possui variáveis continuas, mas para determinar o conjunto de dados frequentes é necessário trabalhar com variáveis discretas.
Esta operação resolve este problema.



Orange 2 – Discretização

4.3. Resultados da avaliação da aplicação do algoritmo Apriori

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



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.247	1.000	0.247	1.351	3.000	0.164	PetalLengthCm=< 1.55	→	Species=Iris-setosa
0.227	1.000	0.227	1.471	3.000	0.151	PetalLengthCm=≥ 5.15	→	Species=Iris-virginica
0.227	1.000	0.227	1.471	3.000	0.151	PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.227	1.000	0.227	1.471	3.000	0.151	PetalWidthCm=≥ 1.85	→	Species=Iris-virginica
0.180	1.000	0.180	1.852	3.000	0.120	PetalLengthCm=< 1.55, PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.173	1.000	0.173	1.923	3.000	0.116	SepalLengthCm=< 5.15, PetalLengthCm=< 1.55	→	Species=Iris-setosa
0.173	1.000	0.173	1.923	3.000	0.116	PetalLengthCm=≥ 5.15, PetalWidthCm=≥ 1.85	→	Species=Iris-virginica
0.167	1.000	0.167	2.000	3.000	0.111	SepalLengthCm=< 5.15, PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.160	0.923	0.173	1.308	4.072	0.121	SepalLengthCm=≥ 6.45, Species=Iris-virginica	→	PetalLengthCm=≥ 5.15
0.160	1.000	0.160	2.083	3.000	0.107	SepalLengthCm=≥ 6.45, PetalLengthCm=≥ 5.15	→	Species=Iris-virginica
0.153	0.920	0.167	1.760	3.136	0.104	PetalLengthCm=1.55 - 4.35, Species=Iris-versicolor	→	PetalWidthCm=0.25 - 1.35
0.140	1.000	0.140	2.381	3.000	0.093	SepalWidthCm=≥ 3.35, PetalLengthCm=< 1.55	→	Species=Iris-setosa
0.140	1.000	0.140	2.381	3.000	0.093	SepalLengthCm=≥ 6.45, PetalWidthCm=≥ 1.85	→	Species=Iris-virginica
0.133	0.909	0.147	1.864	3.326	0.093	PetalWidthCm=1.35 - 1.85, Species=Iris-versicolor	→	PetalLengthCm=4.35 - 5.15
0.127	0.905	0.140	1.619	3.992	0.095	SepalLengthCm=≥ 6.45, PetalWidthCm=≥ 1.85, Species=Iris-virginica	→	PetalLengthCm=≥ 5.15
0.127	0.905	0.140	1.619	3.992	0.095	SepalLengthCm=≥ 6.45, PetalWidthCm=≥ 1.85	→	PetalLengthCm=≥ 5.15
0.127	0.905	0.140	1.619	3.992	0.095	SepalLengthCm=≥ 6.45, PetalWidthCm=≥ 1.85	→	PetalLengthCm=≥ 5.15, Species=Iris-virginica
0.127	1.000	0.127	2.632	3.000	0.084	SepalLengthCm=< 5.15, PetalLengthCm=< 1.55, PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.127	1.000	0.127	2.632	3.000	0.084	SepalLengthCm=≥ 6.45, PetalLengthCm=≥ 5.15, PetalWidthCm=≥ 1.85	→	Species=Iris-virginica
0.113	1.000	0.113	2.941	3.000	0.076	SepalWidthCm=≥ 3.35, PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.113	0.944	0.120	2.778	2.833	0.073	SepalWidthCm=< 2.75, PetalWidthCm=0.25 - 1.35	→	Species=Iris-versicolor
0.107	1.000	0.107	3.125	3.000	0.071	SepalLengthCm=< 5.15, SepalWidthCm=≥ 3.35	→	Species=Iris-setosa
0.107	1.000	0.107	3.125	3.000	0.071	SepalWidthCm=< 2.75, PetalLengthCm=1.55 - 4.35	→	Species=Iris-versicolor
0.100	0.938	0.107	1.750	5.022	0.080	SepalWidthCm=< 2.75, PetalLengthCm=1.55 - 4.35	→	PetalWidthCm=0.25 - 1.35, Species=Iris-versicolor
0.100	0.938	0.107	2.750	3.196	0.069	SepalLengthCm=5.15 - 5.85, PetalLengthCm=1.55 - 4.35, Species=Iris-versicolor	→	PetalWidthCm=0.25 - 1.35
0.100	0.938	0.107	2.750	3.196	0.069	SepalWidthCm=< 2.75, PetalLengthCm=1.55 - 4.35, Species=Iris-versicolor	→	PetalWidthCm=0.25 - 1.35
0.100	0.938	0.107	2.750	3.196	0.069	SepalWidthCm=< 2.75, PetalLengthCm=1.55 - 4.35	→	PetalWidthCm=0.25 - 1.35
0.100	1.000	0.100	3.333	3.000	0.067	SepalWidthCm=< 2.75, PetalLengthCm=1.55 - 4.35, PetalWidthCm=0.25 - 1.35	→	Species=Iris-versicolor
0.100	1.000	0.100	3.333	3.000	0.067	SepalWidthCm=2.75 - 3.05, PetalLengthCm=≥ 5.15	→	Species=Iris-virginica
0.093	1.000	0.093	2.571	4.167	0.071	SepalLengthCm=5.15 - 5.85, Species=Iris-setosa	→	SepalWidthCm=≥ 3.35
0.093	1.000	0.093	3.571	3.000	0.062	SepalLengthCm=5.15 - 5.85, SepalWidthCm=≥ 3.35	→	Species=Iris-setosa
0.087	1.000	0.087	3.846	3.000	0.058	SepalWidthCm=≥ 3.35, PetalLengthCm=< 1.55, PetalWidthCm=< 0.25	→	Species=Iris-setosa
0.087	1.000	0.087	3.846	3.000	0.058	SepalWidthCm=≥ 3.35, PetalWidthCm=0.25 - 1.35	→	Species=Iris-setosa

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.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.247	1.000	0.247	1.351	3.000	0.164	PetalLengthCm=< 1.55	→	Species=Iris-setosa
0.227	1.000	0.227	1.471	3.000	0.151	PetalLengthCm=≥ 5.15	→	Species=Iris-virginica
0.227	1.000	0.227	1.471	3.000	0.151	PetalWidthCm=< 0.25	→	Species=Iris-setosa

Orange 4 - Resultados Apriori (exemplos)

5. 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.

6. 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