DATA MINING - CLASSIFICAÇÂO DAS FLORES DO DATASET IRIS

Cibele Castelo Nogueira

cibele_cast@hotmail.com.br

Pontifícia Universidade Católica de Minas Gerais (PUC Minas- Contagem) – Brasil

Linguagem Estatística – Leonardo Villela

ABSTRACT

This article presents the Data Mining cycle for the development of a predictive analysys problem. It is planned to do the classification of Iris and its respective species throught the implementation of an algorithm KNN (K-Nearest-Neighbor) in the step of Data Modeling. The dataset Iris will be use for the problem solution.

RESUMO

Este artigo apresenta o ciclo de *Data Mining* para o desenvolvimento de um problema de análise preditiva. Pretende-se realizar a classificação das flores e suas respectivas espécies por meio da implementação do algoritmo *KNN (K-Nearest-Neighbor)* na etapa de modelagem do dados. O conjunto de dados Iris será utilizado para a resolução do problema.

PALAVRAS-CHAVE: Data Mining, Iris, *Scikit-learn*, KNN.

1. INTRODUÇÃO

O conjunto de dados IRIS é utilizado para resolução e entendimento de múltiplas medidas de problemas de taxonomia, e serve também para reconhecimento de padrão e estudo de algoritmos de *Data Mining*, *Machine Learning*, *Deep Learning* e inteligência de máquina em geral.

A classificação é uma tarefa que envolve a divisão de objetos para que cada um seja designado para um de um número de categorias exclusivas que são chamadas de classes, assim cada objeto deve ser designado para precisamente uma classe, nunca para mais que uma e nunca para nenhuma classe. (BRAMER, 2016)

A classificação de flores em suas respectivas espécies é um problema de classificação de múltiplas classes, pois envolve dados das características e do tipo de espécie. Os três tipos de espécies possuem características bem diferentes que os tornam diferenciáveis perante uns aos outros. Como os *labels* já estão pré-definidos, será utilizado o algoritmo de classificação KNN (*K-Nearest Neighbors*) para construção do modelo que terá que aprender as medidas das características das flores dessas três espécies para prever qual é a espécie para cada nova flor.

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta o ciclo de data mining e suas etapas iniciais, incluindo também uma análise exploratória. Na Seção 3 são apresentados

o experimento e a análise de resultado e na seção 4 uma conclusão final que avaliará se o modelo construído atende ou não a necessidade do problema de negócio.

2. DESCOBERTA DO CONHECIMENTO E CICLO DATA MINING

Data mining é um processo que envolve várias etapas com o objetivo de resolver um problema por meio da utilização de dados. A figura 1 mostra seu ciclo que envolve as etapas de: Entendimento do problema de negócio; Entendimento dos dados; Preparação dos dados; Construção do modelo; teste e avaliação e entrega.

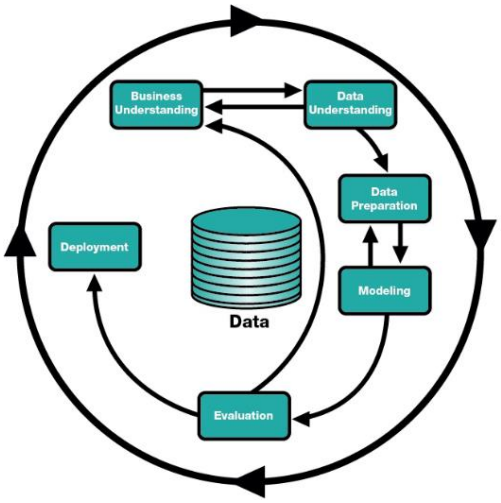

Figura 1: Ciclo de Data Mining
Fonte: Chapman *et al. (*2001)

Existem diversos algoritmos para reconhecimento de padrão em dados. Como por exemplo, algoritmos de clusterização e classificação que são algoritmos que tem como objetivo realizar a divisão dos dados em grupos baseados em características. Estes algoritmos se diferem, pois, algoritmos de classificação já possuem as características pré-definidas para cada *label*, ao contrário de clusterização que pretende descobrir padrões nos dados e agrupa-los de acordo com parâmetros pré-definidos.

2.1 ENTENDIMENTO DO PROBLEMA

Este problema é um problema de classificação de múltiplas classes, pois o conjunto de dados possui características que já estão relacionadas com a variável *target* que é o tipo da espécie de cada flor. Sendo assim, cada flor já possui o Tipo de Espécie e o objetivo é utilizar um algoritmo de classificação que irá predizer qual é o tipo de espécie para novas flores (*que não tiveram suas características no conjunto de dados de treinamento). Para este objetivo, será utilizado um algoritmo de classificação KNN (K-nearest neighbors), que leva em consideração a alteração do parâmetro ‘k números de vizinhos’ para verificação do melhor agrupamento dos dados.

2.2 ENTENDIMENTO DOS DADOS

O conjunto de dados Iris possui a coluna “Tipo de Espécie” e seus atributos: 'setosa', 'versicolor' e 'virginica', contendo cada um 50 amostras de flores, e possui também as colunas de características que são: “comprimento da sépala”, “largura da sépala”, “comprimento da pétala” e “largura da pétala”, totalizando-se assim 5 colunas com um total de 150 linhas. Na figura 2, pode-se observar a localização da Sépala e da Pétala da flor.

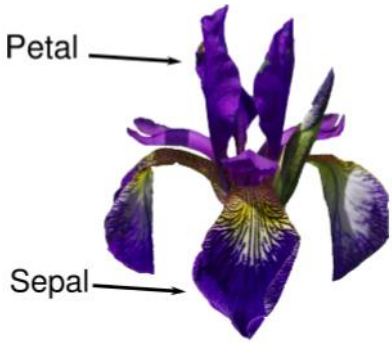

Figura 2: Pétala e Sépala de Iris
Fonte: Guido (2017)

2.2.2 ANÁLISE EXPLORATÓRIA

Na análise exploratória forma utilizados os gráficos de dispersão e de barra que permitem visualizar a distribuição dos dados e ver as relações que existem entre as características e as espécies. Na figura 3, pode-se ver a distribuição de características para cada tipo de espécie. Este gráfico permite concluir por exemplo que: a espécie "virginica" possui Sépalas maiores e a espécie Setosa possui Pétalas maiores.

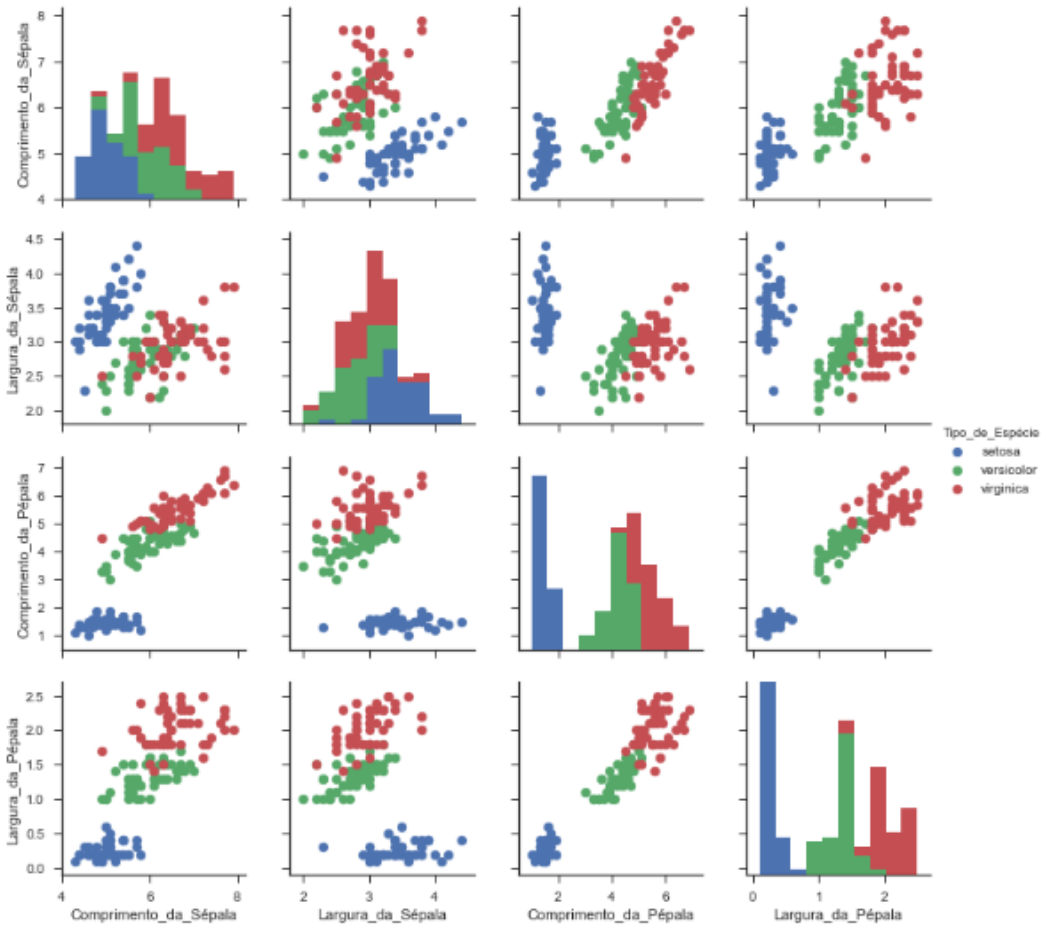

Figura 3: Gráfico de dispersão das características x tipo de espécie
Fonte: Adaptado de Waskon (2017)

2.3 PREPARAÇÂO DOS DADOS

A preparação dos dados tem como objetivo realizar a limpeza e correção dos dados. Esta etapa visa a qualidade dos dados e garante que a base de dados contenha dados precisos e confiáveis. Sua importância é muito grande, pois senão for feita, as próximas etapas podem ser prejudicadas pois conterão dados sujos, duplicados e valores faltantes que irão impactar diretamente na avaliação final do modelo e consequentemente na consistência da informação. Para isso, utilizou-se a biblioteca Pandas que fornece suporte para estruturação e análise de dados. Ela possui familiaridade com a linguagem de consulta estruturada SQL e permite fazer operações DML e DLL dentro de sua própria sintaxe.

Realizou-se análise, extração e transformação do *dataset* Iris. Entre as tarefas executadas, cito: junção dos dataframes de *"data"* e *"tipo_de_especie* ; Substituição de valores NAN por 0 para o Tipo de Espécie e alteração dos nomes das colunas.

Assim sendo, a manipulação de dados permitiu obter informações dos conjuntos de dados, possibilitando verificar, selecionar e atualizar campos que haviam necessidade de alteração.

A figura 4, mostra a visualização do *DataFrame* que irá ser utilizado para as etapas posteriores.

	Comprimento da Pétala	Comprimento da Sépala	Largura da Pétala	Largura da Sépala	Tipo da Espécie
0	1.4	5.1	0.2	3.5	0
1	1.4	4.9	0.2	3	0
2	1.3	4.7	0.2	3.2	0
3	1.5	4.6	0.2	3.1	0
4	1.4	5	0.2	3.6	0
5	1.7	5.4	0.4	3.9	0
6	1.4	4.6	0.3	3.4	0
7	1.5	5	0.2	3.4	0
8	1.4	4.4	0.2	2.9	0
9	1.5	4.9	0.1	3.1	0

Figura 4: *DataFrame* com os dados de Iris
Fonte: Autoria própria.

3.EXPERIMENTO E RESULTADO

3.1.MODELAGEM

A modelagem dos dados será feita com a utilização da biblioteca *Scikit-learn* e do algoritmo escolhido KNN (K-nearest neighbors), que leva em consideração a alteração do parâmetro K ("k pontos vizinhos dos dados de treinamento") para verificação do melhor agrupamento dos dados. Este algoritmo realiza a previsão de um novo "ponto de dados", ele encontra o ponto no conjunto de dados de treinamento que é mais próximo desse novo ponto, e então ele o designa para esse rótulo/*label* desse ponto de treinamento.

Na figura 5 pode-se ver o *label* Tipo de Espécie que será o *target*, assim sendo cada flor, com suas características, possui um tipo de espécie.

	Tipo de Espécies
0	setosa
1	versicolor
2	virginica

Figura 5: *DataFrame* com o tipo de espécies
Fonte: Autoria própria.

Dessa forma, esse modelo utilizou-se de todas as quatro colunas de características: “comprimento da sépala”, “largura da sépala”, “comprimento da pétala” e “largura da pétala” como as *features*/características e como variável *target*: “Tipo_de_Especie”

O modelo de dados foi construído com 75% por cento de dados de treinamento e 25% por cento de dados para teste. Divide-se o *dataset* em conjunto de dados de treino para construir o modelo, e o outro conjunto de dados de teste para avaliar se o modelo terá um bom desempenho com novos dados (*que não foram vistos antes). Desse modo, temos o benefício de se ter dados de teste para avaliação do desempenho do modelo para evitar, por exemplo, o problema de overfitting dos dados. Se não houver esta divisão dos *datasets*, o modelo pode não ter precisão e passar a ser inadequado para a determinada solução, pois não seria capaz de prever novos dados, e por isso não iria atender ao problema de negócio.

3.2 AVALIAÇÃO DO MODELO

A avaliação do modelo é uma etapa importante para saber se o modelo cumpriu com objetivo. Por esse motivo, como falado anteriormente, o conjunto de dados foi dividido em dados de treinamento e dados de teste. Os dados de teste são aqueles que não foram vistos pelo modelo e serão utilizados nesta etapa para verificar se o modelo consegue generalizar para dados não vistos.

Para realização dessa etapa, será utilizado o coeficiente de determinação R^2 como métrica de avaliação da classificação do *dataset* IRIS. O coeficiente de determinação r^2 foi utilizado para quantificar o desempenho do modelo, ou seja, mostrar o quão bom é o modelo em fazer predições. Os valores variam de 0 até 1, quanto mais próximo de 1 melhor será a predição do valor da variável target.

O resultado obtido para o conjunto de teste foi de 0.97% e para o conjunto de treinamento de 100%.

4.CONCLUSÃO

Pretendeu-se com esse artigo apresentar um algoritmo de data mining para reconhecimento de padrão e análise preditiva de classificação das flores do conjunto de dados IRIS, por meio da estruturação de ciclo de data mining, com o intuito de aprofundamento no entendimento de cada etapa para entendimento do problema proposto.

O algoritmo KNN utilizado, possibilitou a obtenção de um resultado de 97,0% de precisão, o que demonstra a eficácia do modelo na generalização de dados ainda não vistos.

Conclui-se assim que, modelo construído atingiu o seu propósito de classificação das flores em suas respectivas espécies.

REFERÊNCIAS

CHAPMAN, Pete *et al.* **Phases of the CRISP-DM reference model.**2001.Disponível em: <http://crisp-dm.eu/>. Acesso em: 05 de jun. 2017.
FISHER,R.A. **Iris DataSet.** Disponível em: <https://archive.ics.uci.edu/ml/datasets/Iris>. Acesso em: 07 de jun. 2017.
WASKOM, Michael**. Visualizing the distribution of a dataset.** 2012-2015. Disponível em: <http://seaborn.pydata.org/tutorial/axis_grids.html>. Acesso em: 10 de jun. 2017.

BIBLIOGRAFIA CONSULTADA

BRAMER, Max. **Principles of Data Mining**. London, New Jersey,Springer, pp.520, 2016.
BROWN, Meta S. Brown. **Data Mining for Dummies**. John Wiley & Sons, Inc, pp.408, 2014.
GUIDO, S. and Müller, A**. Introduction to Machine Learning with Python**. United States of America, O'Reilly Media, pp.376, 2017.
MCKINNEY, Wes. **Python for Data Analysis**. United States of America, O'Reilly Media, pp.451, 2013.