TRABALHO DE SISTEMAS INTELIGENTES
Pratica de Conceitos Aprendidos na disciplina

Alice Penido, Guilherme Oliveira
Pontífícia Universidade Católica de Minas Gerais

4 de dezembro de 2023

Resumo

Este artigo tem como objetivo apresentar os resultados da prática de conceitos relacionados ao aprendizado de máquina (Machine Learning) utilizando ferramentas tecnológicas. O projeto desenvolvido consistiu em obter dados, prepará-los e aplicar três modelos de aprendizado de máquina, possibilitando assim a previsão de valores relacionados às bases de dados obtidas. Os experimentos foram realizados com o uso da ferramenta Orange Data Mining.

Palavras-chave: Machine Learning, Orange Software, Orange Data Mining, Random Forest, Neural Network, Decision Tree.

1 Introdução

O campo do Machine Learning experimentou uma notável popularização, principalmente na última década, impulsionada pela descoberta e aprimoramento de fórmulas matemáticas inovadoras, juntamente com a crescente necessidade de poder descobrir o que acontecerá, com base em dados e estruturas já existentes.

Neste estudo, nos propusemos a achar uma base de dados de interesses, tratando valores considerados outliers, selecionando colunas relevantes e utilizando modelos de machine learning para predição, analisando suas métricas de acerto, aplicando todos os conhecimentos adquiridos durante o curso referente com auxílio do Orange.

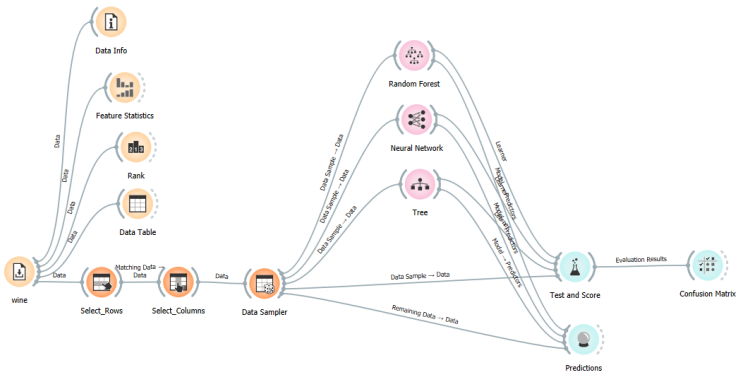


Figura 1: Fluxo

2 Dados

Para uso neste artigo, escolhemos utilizar uma das base já disponibilizadas no Orange, nomeada Wine.

	Wine	Alcohol	Color intensity	)/OD315 of diluted	Proline	Flavanoids	Total phenols	Hue
60	2	11.61	2.65	3.26	680	2.92	2.74	0.960
47	2	11.64	2.8	2.75	680	1.69	1.95	1.000
41	2	11.66	3.8	2.14	428	1.57	1.61	1.230
55	2	11.82	2.06	2.44	415	1.64	2.50	0.940
42	2	11.84	2.65	2.52	500	1.32	1.72	0.960
45	2	11.84	3.05	3.08	520	2.21	2.20	0.790
49	2	12.00	3.6	2.65	450	1.25	1.45	1.050
67	2	12.04	2.6	2.57	580	1.75	2.10	0.790

Figura 2: Exemplo da base

A base continha 178 linhas e 14 colunas, não contendo valores vazios. Além da variável a ser descoberta - a classe do vinho - nenhuma outra variável era categórica, sendo divisível em dois subgrupos:

1 - Continuas: Alcohol, Malicacid, Ash, Alcalinity_of_ash, Total_phenols, Flavanoids, Nonflavanoid_phenols, Proanthocyanins, Color_intensity, Hue, 0D280_0D315_of_diluted_wines 2 - Inteiras: Magnesium, Proline

Como pode ser observado, a base em sua maioria utiliza de números contínuos. Não vimos necessidade de normalização ou discretização de dados, a partir do node Features Statistics.

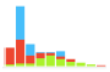
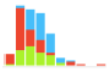
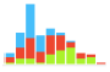
	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	Alcohol		13.0006	12.37	13.05	0.0623	11.03	14.83	0 (0 %)
N	Malic Acid		2.3363	1.73	1.8650	0.4768	0.74	5.80	0 (0 %)
N	Ash		2.3665	2.28	2.36	0.1156	1.36	3.23	0 (0 %)
N	Alcalinity of ash		19.495	20.0	19.5	0.171	10.6	30.0	0 (0 %)
N	Magnesium		99.74	88	98	0.14	70	162	0 (0 %)
N	Total phenols		2.2951	2.20	2.3550	0.2719	0.98	3.88	0 (0 %)
N	Flavanoids		2.0293	2.65	2.1350	0.4908	0.34	5.08	0 (0 %)
N	Nonflavanoid phenols		0.3619	0.26	0.34	0.3430	0.13	0.66	0 (0 %)
N	Proanthocyanins		1.5909	1.35	1.5550	0.3588	0.41	3.58	0 (0 %)
N	Color intensity		5.05809	2.6	4.69	0.457043	1.28	13	0 (0 %)
N	Hue		0.95745	1.040	0.965	0.23806	0.480	1.710	0 (0 %)
N	OD280/OD315 of diluted wines		2.6117	2.87	2.78	0.2711	1.27	4.00	0 (0 %)

Figura 3: Descrição dos dados

3 Pré-processamento

A solução de problemas baseada em dados requer dados bem estruturados, consistentes e relevantes. Portanto, a partir dos dados já disponibilizados para aprendizado dentro do próprio Orange e do node de Box Plot, fizemos a análise de outliers e colunas relevante para aprendizado.

Foi decidido pela equipe a retirada de dados que, após representados por meio de um node Box Plot, visualmente não estejam entre o primeiro e o terceiro quartis em todas variáveis, tendo assim tomado por removido os outliers da base de dados.

Logo após, utilizamos do node de Rank para decidir quais colunas utilizariamos, tendo visto necessidade apenas no top 7, que tinham tanto Gain Ratio quanto Gini como valores vistos como relevantes.

		#	Gain ratio	Gini
1	N Flavanoids		0.429	0.339
2	N Proline		0.391	0.311
3	N Color intensity		0.332	0.265
4	N OD280/OD315 of diluted wines		0.321	0.243
5	N Alcohol		0.294	0.250
6	N Hue		0.293	0.213
7	N Total phenols		0.270	0.193
8	N Proanthocyanins		0.169	0.120
9	N Alcalinity of ash		0.165	0.120
10	N Malic Acid		0.154	0.140
11	N Magnesium		0.143	0.122
12	N Nonflavanoid phenols		0.119	0.095
13	N Ash		0.054	0.053

Figura 4: Ranqueamento de variáveis relevantes

4 Metodologia

O desenvolvimento do trabalho consistiu na implementação de três modelos supervisionados par prever a classificação de classes de vinho, todos utilizando-se de 70% dos dados para treino e 30% para teste. A seguir uma breve descrição desses algoritmos.

4.1 Decision Tree

A Decision Tree é um método supervisionado que constrói modelos de classificação ou regressão em uma estrutura semelhante a uma árvore. Foi estabelecido pela primeira vez em 1963 por Morgan e Sonquist. Esse método é:
(1) conceitualmente simples, mas poderoso;
(2) intuitivo para interpretação;
(3) capaz de lidar com valores ausentes e características mistas; e
(4) capaz de selecionar variáveis automaticamente.
No entanto, seu poder preditivo não é excessivamente competitivo. Geralmente, a árvore de decisão não é estável com alta variância do modelo e pequenas variações nos dados de entrada resultariam em um grande efeito na estrutura da árvore.

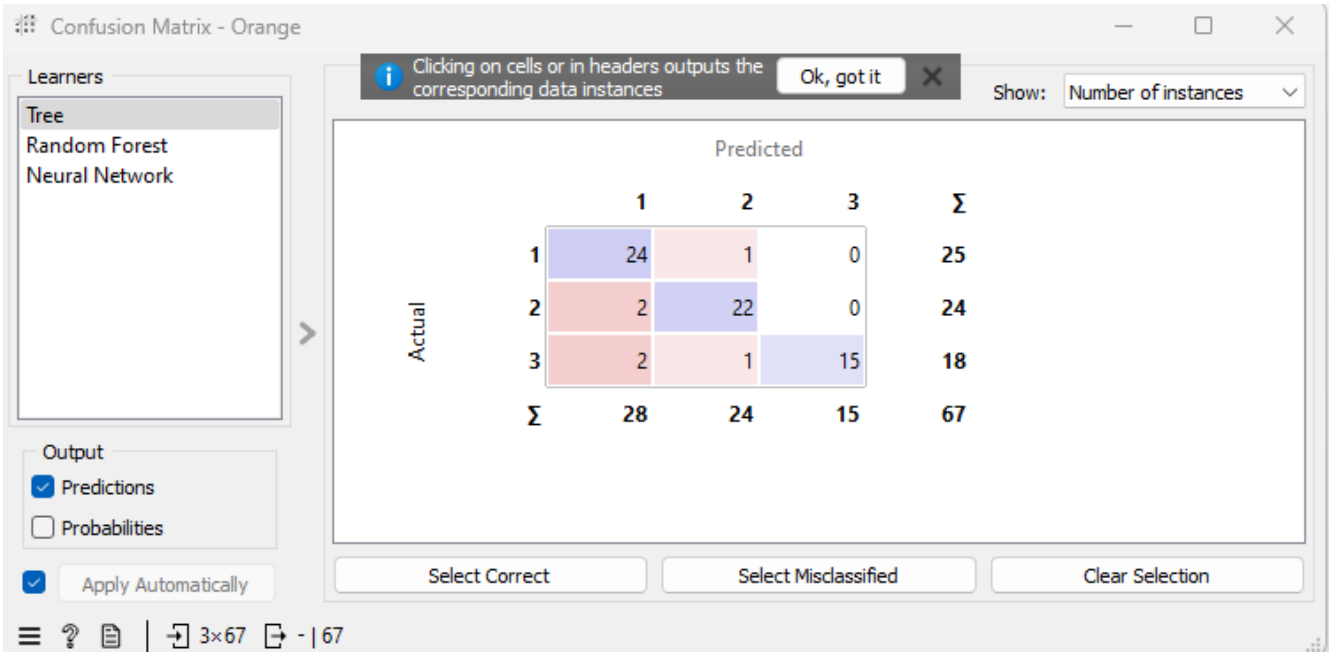


Figura 5: Matriz de confusão da Decision Tree

4.2 Random Forest

Random Forests utilizam um método de conjunto que melhora a estrutura básica da árvore de decisão ao combinar diversos aprendizes fracos para criar um aprendiz mais robusto. Nesse método, várias árvores de decisão (os aprendizes fracos) são construídas com conjuntos de treinamento bootstrap. Para cada árvore de decisão, é escolhida uma amostra aleatória de "m"preditores dos "P"preditores totais. Como "m"é menor que "P", a maioria dos preditores não é considerada, o que evita que algumas características tenham grande influência sobre todas as árvores individuais. Ao combinar essas árvores não correlacionadas por meio da média, é possível reduzir a variância, tornando o resultado final mais estável e confiável.

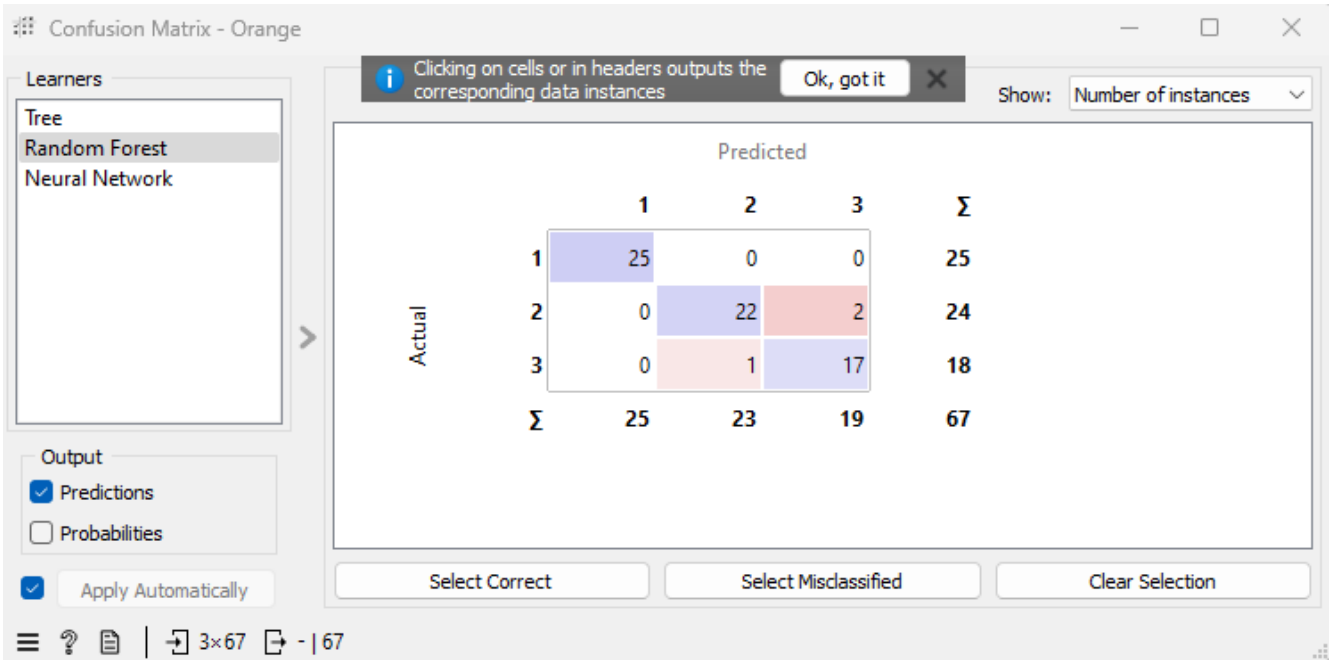


Figura 6: Matriz de confusão da Random Forest

4.3 Neural Network

Redes neurais, também conhecidas como perceptrons de múltiplas camadas, são projetadas para simular as operações do sistema nervoso humano. A forma mais simples de uma rede neural é um único perceptron. Os elementos essenciais para um perceptron incluem valores de entrada, pesos associados, viés, funções de ativação e uma saída computada. Funções de ativação comumente usadas incluem a sigmoide, tangente hiperbólica (Tanh) e unidades lineares retificadas (ReLU). Uma rede neural pode conter mais de uma camada entre a entrada e a saída para lidar com problemas complexos.
Nosso estudo se utilizou de 100 neurônios artificiais por camada, para produzir os resultados.

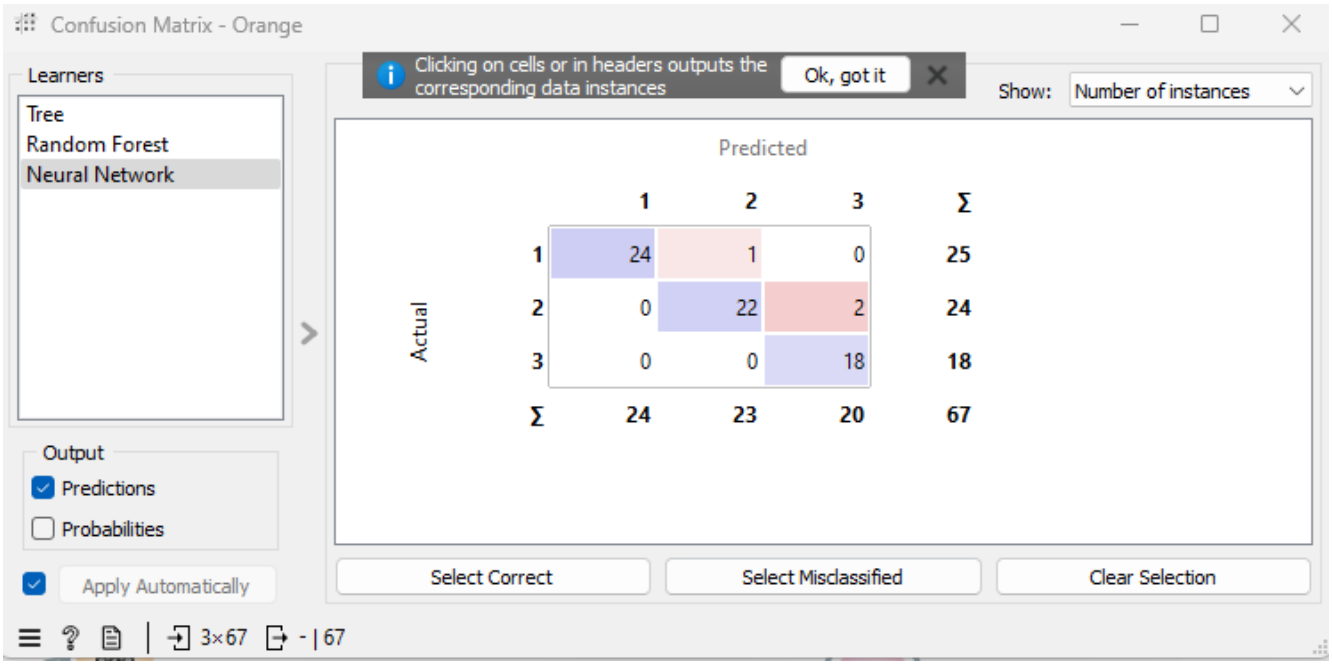


Figura 7: Matriz de confusão da Neural Network

5 Resultados e Considerações Finais

O trabalho resultou na comparação dos resultados dos três modelos, sendo visível que, para esses problema, representado com esses dados e via esse tratamento, o melhor modelo foi o Random Forest.

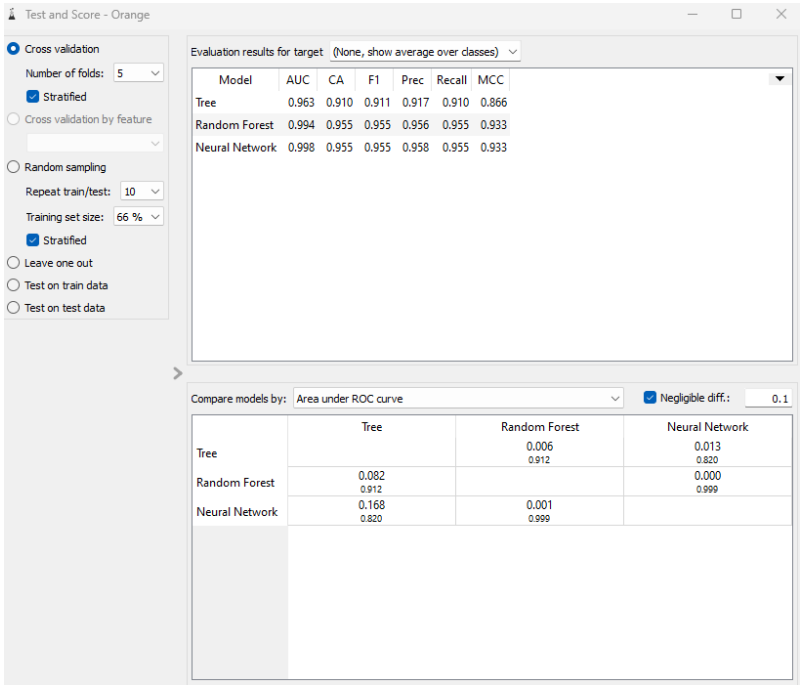


Figura 8: Comparação entre modelos

Pedindo também para os modelos preverem quais seriam as prováveis classes de vinho, podemos ver que todos tendem a acertar os dados, provando assim que tanto a coleta quanto o tratamento de dados foram eficientes.

		error	Random Forest	error	Tree	error	Wine
1	3	0.066	3	0.188	3	0.000	3
2	2	0.002	2	0.183	3	0.750	2
3	2	0.003	2	0.125	2	0.000	2
4	2	0.002	2	0.417	3	0.750	2
5	3	0.001	3	0.000	3	0.000	3
6	2	0.025	2	0.000	2	0.000	2
7	1	0.001	1	0.000	1	0.000	1
8	3	0.003	3	0.188	3	0.000	3
9	2	0.020	2	0.254	2	0.000	2
10	2	0.447	2	0.400	3	1.000	2
11	3	0.002	3	0.075	3	0.000	3
12	2	0.464	2	0.475	3	1.000	2
13	1	0.007	1	0.000	1	0.000	1
14	1	0.001	1	0.000	1	0.000	1
15	2	0.004	2	0.300	3	1.000	2
16	1	0.349	1	0.600	1	0.500	1
17	1	0.001	1	0.000	1	0.000	1
18	1	0.003	1	0.100	1	0.000	1
19	3	0.008	3	0.000	3	0.000	3
20	3	0.006	3	0.300	3	0.000	3

Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
	1.000	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
Tree	0.913	0.821	0.810	0.881	0.821	0.769

Figura 9: Previsão para a base de testes

6 Referências Bibliográficas

1. ZHAO, Yue et al. Employee turnover prediction with machine learning: A reliable approach. In: Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2. Springer International Publishing, 2019. p. 737-758.

2. UCI Machine Learning Repository. (1991). Wine Data Set. Obtido de: https://archive.ics.uci.edu/dataset/109/wine