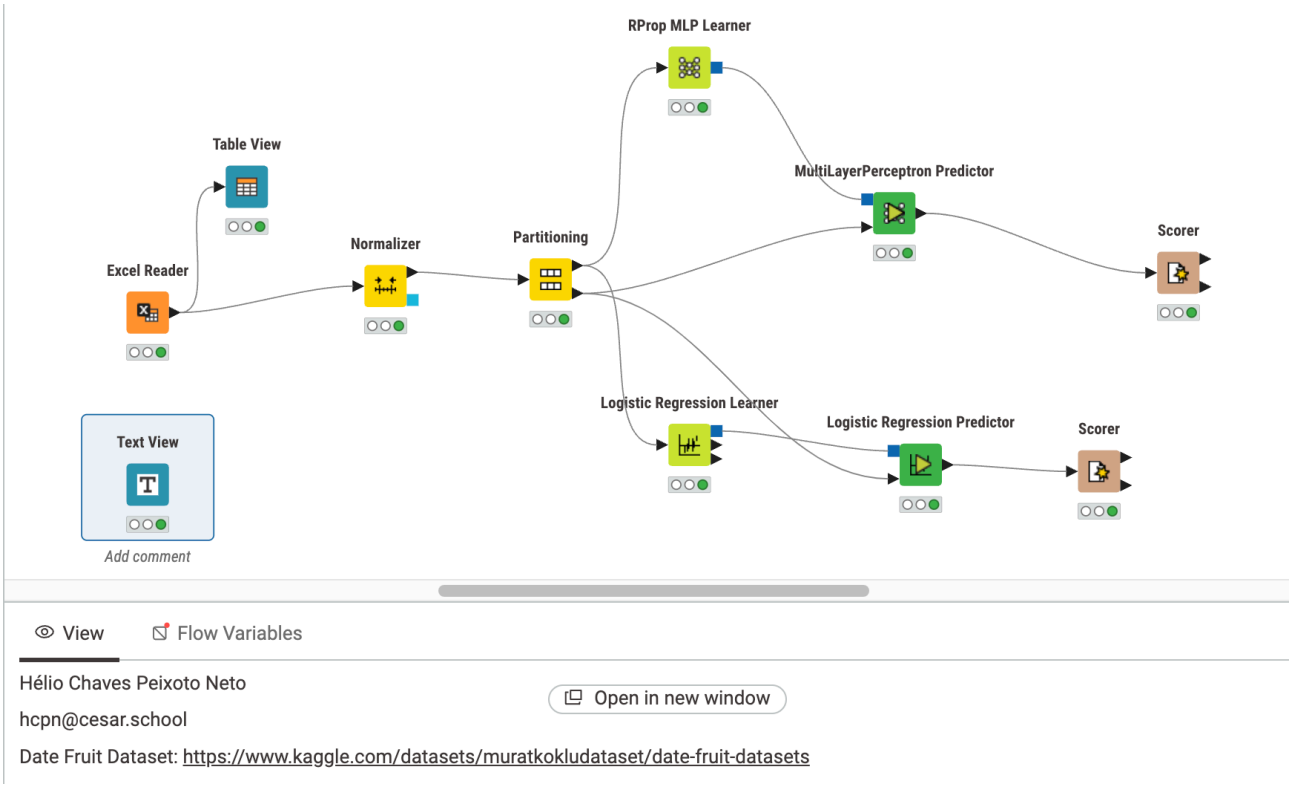
TRABALHO FINAL

Hélio Chaves Peixoto Neto

hcpn@cesar.school

Date Fruit Dataset: https://www.kaggle.com/datasets/muratkokludataset/date-fruit-datasets



Contexto e Análise

Meu objetivo com o *workflow* mostrado acima foi comparar um modelo que utiliza múltiplas camadas de Redes Neurais (o *MultiLayerPerceptron - MLP* -, que nesse caso foi treinado com 3 camadas escondidas, 7 neurônios por camada e 2000 épocas) com um modelo tradicional de *Machine Learning*, a Regressão Logística. Como a escolha do modelo tradicional sugere, o problema do dataset Date Fruit é de classificação, então baseados em 34 parâmetros o algoritmo deve dizer qual a classe de Tâmara está relacionada àquelas características.

Depois de carregar o dataset e visualizar a tabela de dados como um todo, o intuito era fazer um treinamento do modelo, apenas com normalização, para verificar como a predição se comportava. Caso o modelo não tivesse uma boa performance, a ideia era retirar algumas variáveis que pudessem estar provocando algum '*noise*' no resultado.

Para minha surpresa, depois de variar o número de épocas, camadas escondidas e neurônios, o modelo respondeu muito bem, com uma acurácia geral de 91.7%, apesar da classe 'SOGAY' ter uma acurácia de aproximadamente 70%, o que pode ser explicada por sua baixa

representatividade (16 amostras) em relação a classes como ‘SAFAVI’, ‘ROTANA' e 'DOKOL' (cada uma com pelo menos 35 amostras). As imagens abaixo mostram a matriz de confusão e a acurácia do modelo com uso de MLP.

#	RowID	BERHI Number (integ...	DEGLET Number (integ...	DOKOL Number (integ...	IRAQI Number (integ...	ROTANA Number (integ...	SAFAVI Number (integ...	SOGAY Number (integ...
1	BERHI	14	0	0	0	2	0	0
2	DEGLE	0	12	0	0	0	0	3
3	DOKOL	0	2	39	0	0	0	0
4	IRAQI	0	0	0	12	0	1	0
5	ROTAN	0	0	0	0	34	0	1
6	SAFAVI	0	0	0	0	0	43	1
7	SOGAY	1	1	0	1	2	0	11

Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
0.875	0.933	0.875	0.994	0.903	?
0.8	0.8	0.8	0.982	0.8	?
0.951	1	0.951	1	0.975	?
0.923	0.923	0.923	0.994	0.923	?
0.971	0.895	0.971	0.972	0.932	?
0.977	0.977	0.977	0.993	0.977	?
0.688	0.688	0.688	0.97	0.688	?
?	?	?	?	?	0.917

Como aprendemos na disciplina, o uso de várias camadas de Redes Neurais podem resolver problemas mais complexos, então decidi testar o mesmo caso de uso com a Regressão Logística, para entender se o algoritmo responderia bem ou uma etapa de pré-processamento seria essencial. Novamente para a minha surpresa, dada a quantidade de parâmetros envolvidos, o modelo também respondeu bem, obtendo a mesma acurácia de 91.7% do treinamento com MLP. Da mesma forma, a classe ‘DEGLE' está com uma acurácia baixa de 60%, mas possui uma representatividade baixa (16 amostras). Abaixo é mostrada a matriz de confusão e as métricas desse treinamento.

#	RowID	BERHI Number (integ...	DEGLET Number (integ...	DOKOL Number (integ...	IRAQI Number (integ...	ROTANA Number (integ...	SAFAVI Number (integ...	SOGAY Number (integ...
1	BERHI	15	0	0	1	0	0	0
2	DEGLE	0	9	3	0	1	0	2
3	DOKOL	0	2	39	0	0	0	0
4	IRAQI	0	0	0	12	1	0	0
5	ROTAN	1	0	0	0	33	0	1
6	SAFAVI	0	0	0	1	0	43	0
7	SOGAY	0	0	0	0	1	1	14

Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
0.938	0.938	0.938	0.994	0.938	
0.6	0.818	0.6	0.988	0.692	
0.951	0.929	0.951	0.978	0.94	
0.923	0.857	0.923	0.988	0.889	
0.943	0.917	0.943	0.979	0.93	
0.977	0.977	0.977	0.993	0.977	
0.875	0.824	0.875	0.982	0.848	
					0.917

Conclusão

Apesar da mesma acurácia ser atingida nos dois casos, o modelo de Regressão Logística utiliza menos processamento computacional em relação ao MLP. Por isso, há uma tendência de indicação do uso de Regressão neste caso, mas observo dois cuidados importantes para tomar a decisão. O primeiro, não foi feito pré-processamento dos dados; talvez após essa etapa o modelo de MLP possa ter uma performance superior, que só se justificaria seu uso caso o contexto pedisse uma acurácia muito próxima dos 100%, já que a obtida foi muito alta. O segundo, a baixa acurácia da classe ‘DEGLE' no uso da Regressão Logística. Se em um caso prática essa classe tiver mais importância que as demais, por exemplo, o modelo de MLP pode ser uma alternativa interessante.

Como próximos passos para esse problema, eu adicionaria uma robusta etapa de pré-processamento, para entender quais parâmetros mais impactam na predição e realizar alguns treinamentos excluindo os parâmetros menos significativos.