C.E.S.A.R school

Trabalho Final - Data Science e IA

2025 | JANEIRO

Alunos:
Daniel Souza Affonso Ferreira - dsaf@cesar.school
Gustavo Lima Guerreiro - glg@cesar.school

Turma:

Especialização em Tech Lead - 2024.2

SOBRE O DATASET

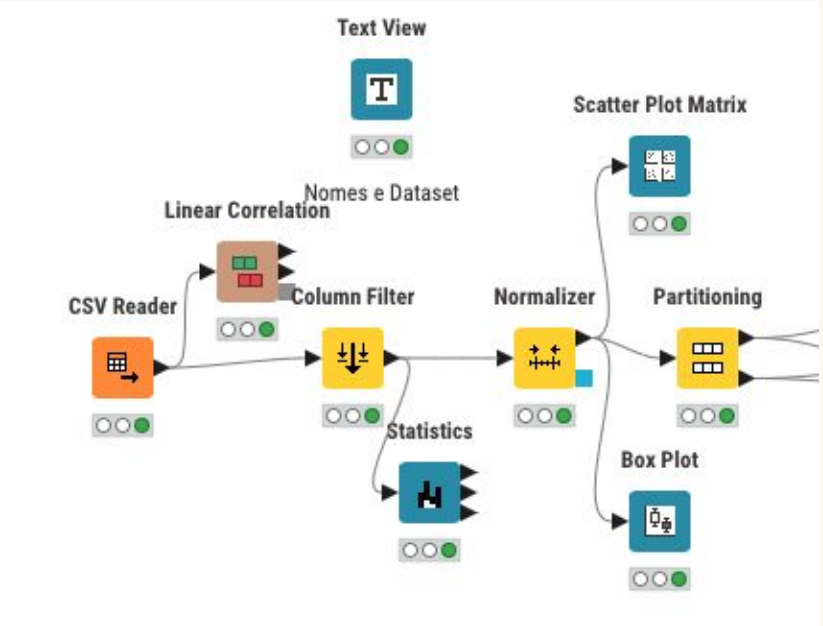
O dataset escolhido foi o Bank Loan Approval, disponível no Kaggle. Nós optamos por utilizá-lo pois estamos desenvolvendo, por meio de nossa empresa, um sistema para o meio financeiro e achamos que seria uma boa oportunidade de nos aprofundar no conteúdo.

Durante a análise inicial, nós observamos a presença de 14 colunas no dataset e 5000 linhas, mas após importá-lo e utilizar uma correlação linear, observamos baixa correlação com a variável observada (*PersonalLoan*), nas seguintes colunas: *ID*, *Age*, *Experience*, *Securities.Account*, *Online*, *Creditcard*. Todas elas tinham correlação abaixo de 0.03, por isso, as descartamos para melhorar a eficiência do nosso modelo.

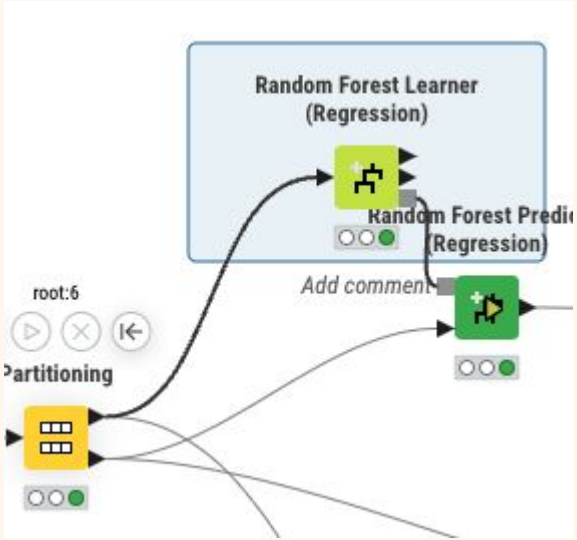
SOBRE OS MODELOS TREINADOS

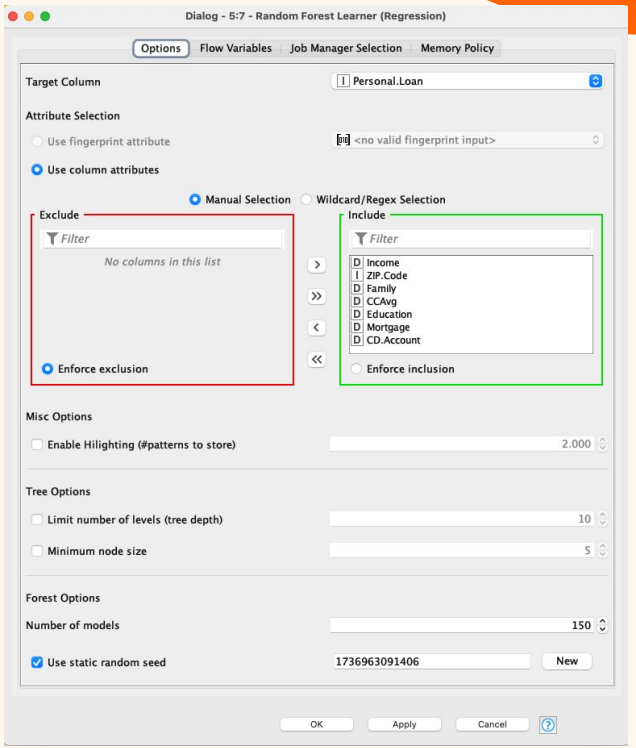
Para uma melhor resolução do problema, nós usamos dois tipos de regressão: A Random Forest e a Regressão Logísticas. No fim, utilizamos um scorer e comparamos o resultado das duas regressões para optar pela que tivesse maior assertividade no nosso caso. As duas regressões apresentaram uma precisão de mais de 90%, nos deixando bem satisfeito com o resultado.

CARREGAMENTO DO DATASET

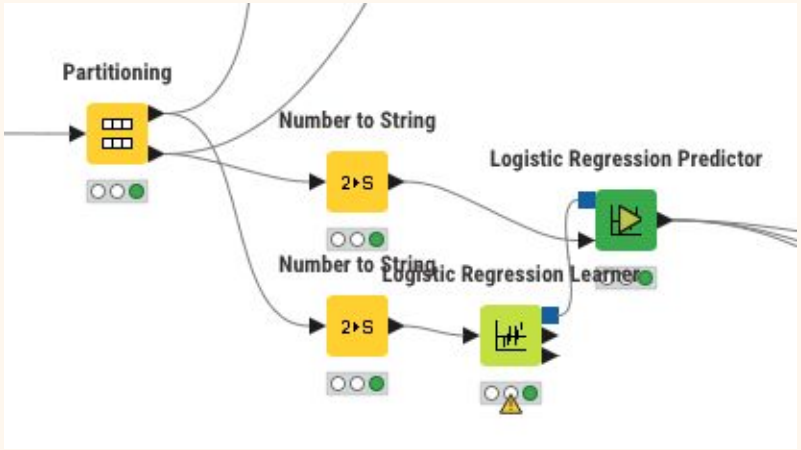


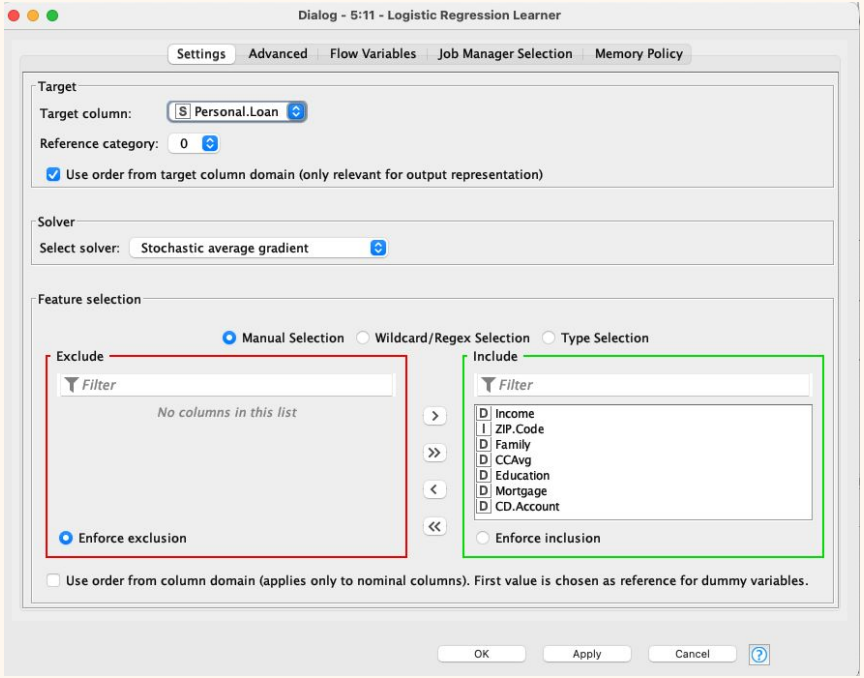
DEFINIÇÃO DE MODELO DA REDE - RANDOM FOREST



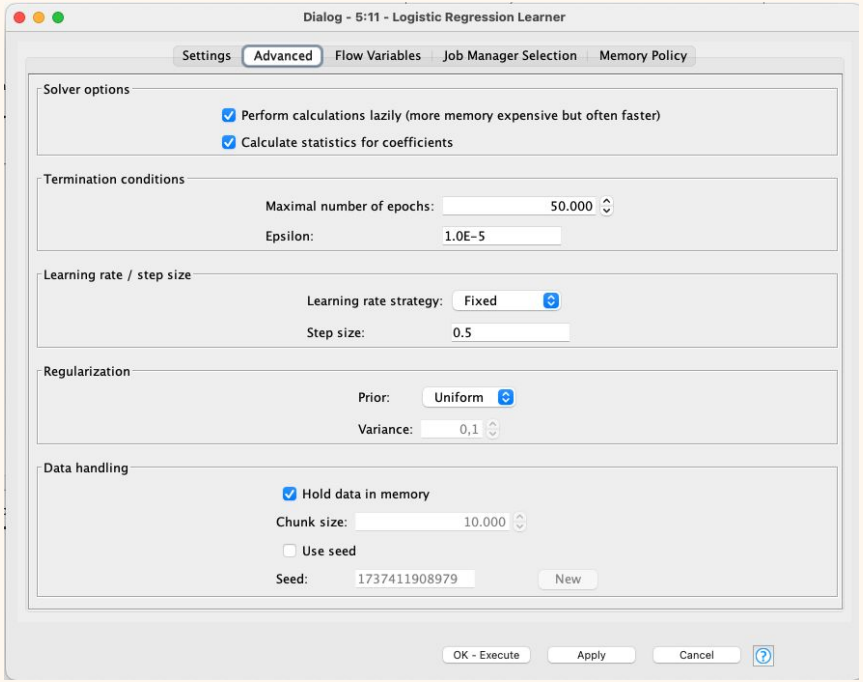


DEFINIÇÃO DE MODELO DA REDE - REGRESSÃO LOGÍSTICA

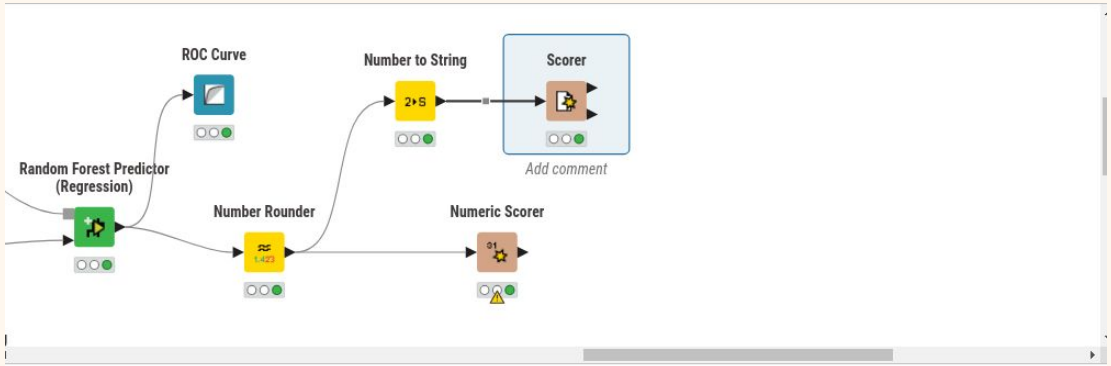




DEFINIÇÃO DE MODELO DA REDE - REGRESSÃO LOGÍSTICA



TREINAMENTO DO MODELO - RANDOM FOREST



▶ 1: Confusion matrix ▶ 2: Accuracy statistics Flow Variables

Rows: 3 | Columns: 11 Table Statistics

	#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-me Numbe
☐	1	0	1125	2	106	17	0.985	0.998	0.985	0.981	0.992
☐	2	1	106	17	1125	2	0.981	0.862	0.981	0.985	0.918
☐	3	Overall	?	?	?	?	?	?	?	?	?

TREINAMENTO DO MODELO - REGRESSÃO LOGÍSTICA

▶ 1: Confusion matrix ▶ 2: Accuracy statistics Flow Variables

Rows: 3 | Columns: 11 Table Statistics

#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-me Numbe
1	0	1127	123	0	0	1	0.902	1	0	0.948
2	1	0	0	1127	123	0	?	0	1	?
3	Overall	?	?	?	?	?	?	?	?	?

CONCLUSÃO

Após a realização de testes com dois modelos de regressão treinados no mesmo conjunto de dados, observou-se que o modelo de *random forest* alcançou uma precisão de 99,8%, enquanto a regressão logística obteve 90,2%. Além de apresentar desempenho superior, o *random forest* demonstrou ser mais eficiente em termos computacionais, atingindo resultados expressivos sem a necessidade de grande número de iterações. Em contrapartida, a regressão logística exigiu 50 mil épocas para alcançar a convergência dos dados.

Ambos os modelos apresentaram taxas de acerto superiores a 90%, evidenciando a qualidade e consistência do dataset utilizado. Esse alto desempenho geral sugere que o problema analisado é relativamente bem estruturado e favorável à solução por meio de métodos de aprendizado de máquina. Assim, o *random forest* não apenas se destaca como a escolha mais eficiente e precisa para o problema em questão, mas também reforça a importância de balancear performance e custo computacional na escolha do modelo ideal.