RELATÓRIO FINAL – IFSP: Orange

Henrique Carvalho de Andrade

Telefone: +55 (12) 99706-2006

E-mail: rique9906@gmail.com

Instituição: Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP

Endereço: R. Antônio Fogaça de Almeida, 200 - Jardim America, Jacareí - SP, 12322-030

CEP: 12322-030

Telefone: (12) 2128-5200

E-mail: prp@ifsp.edu.br

Orientadora: Ana Paula

Telefone: _____

E-mail: _____

RESUMO

Este relatório apresenta o processo de treinamento de uma rede neural no ambiente Orange, utilizando o dataset Iris. Inicialmente, foi feito o treinamento com todos os dados disponíveis. Posteriormente, realizou-se uma análise exploratória para identificar e excluir instâncias conflitantes ou outliers. Após essa etapa de pré-processamento, o modelo foi novamente treinado e seus resultados comparados. O estudo evidencia a importância do tratamento de dados para a obtenção de modelos de classificação mais robustos e confiáveis.

APRESENTAÇÃO

Introdução

O presente estudo trata da aplicação de técnicas de aprendizado de máquina em um conjunto de dados clássico (Iris), por meio da plataforma Orange. O objetivo foi compreender o impacto do pré-processamento dos dados e da remoção de instâncias problemáticas no desempenho de um modelo de classificação baseado em redes neurais artificiais.

Justificativa

A justificativa para este trabalho está na relevância de compreender como dados conflitantes ou fora do padrão podem comprometer o desempenho de modelos de machine learning. Ao tratar corretamente esses dados, é possível alcançar maior precisão e confiabilidade nos resultados, aspecto fundamental para aplicações reais e em larga escala.

Objetivos

O objetivo geral deste estudo é analisar, comparar e discutir sobre o impacto do pré-processamento de dados sobre o desempenho de um modelo de classificação no Orange.

Objetivos específicos:

- Treinar uma rede neural com todos os dados disponíveis do dataset Iris;
- Realizar análise exploratória dos resultados iniciais;
- Identificar e excluir instâncias conflitantes/outliers;
- Re-treinar o modelo com os dados refinados;
- Comparar os resultados obtidos antes e depois do tratamento.

DESENVOLVIMENTO

Metodologia

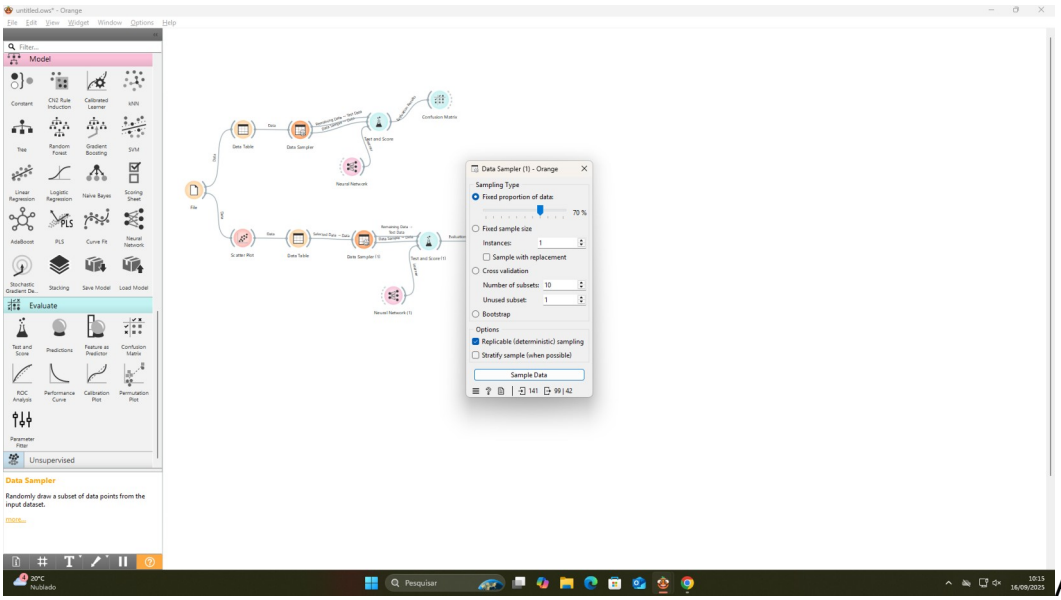
Foi utilizado o dataset Iris, disponível na plataforma Orange, composto por 150 amostras de flores divididas em três espécies. Os atributos empregados foram comprimento e largura da sépala, e comprimento e largura da pétala. No pré-processamento, os atributos numéricos foram normalizados e foram utilizados métodos para detecção de instâncias problemáticas, como análise de outliers (IQR, z-score), conflitos de classificação via k-NN e isolamento de anomalias por Isolation Forest.

O modelo escolhido foi uma Rede Neural Multilayer Perceptron (MLP), configurada com duas camadas ocultas de 10 neurônios cada, função de ativação ReLU, solver Adam e um parâmetro de regularização (alpha) de 0.0001.

O estudo foi dividido em duas fases de treinamento:

1. **Treinamento inicial**: O conjunto de dados foi dividido de forma estratificada, utilizando 70% das amostras para treinamento e 30% para teste. Nesta etapa, o modelo foi treinado com todos os dados disponíveis.

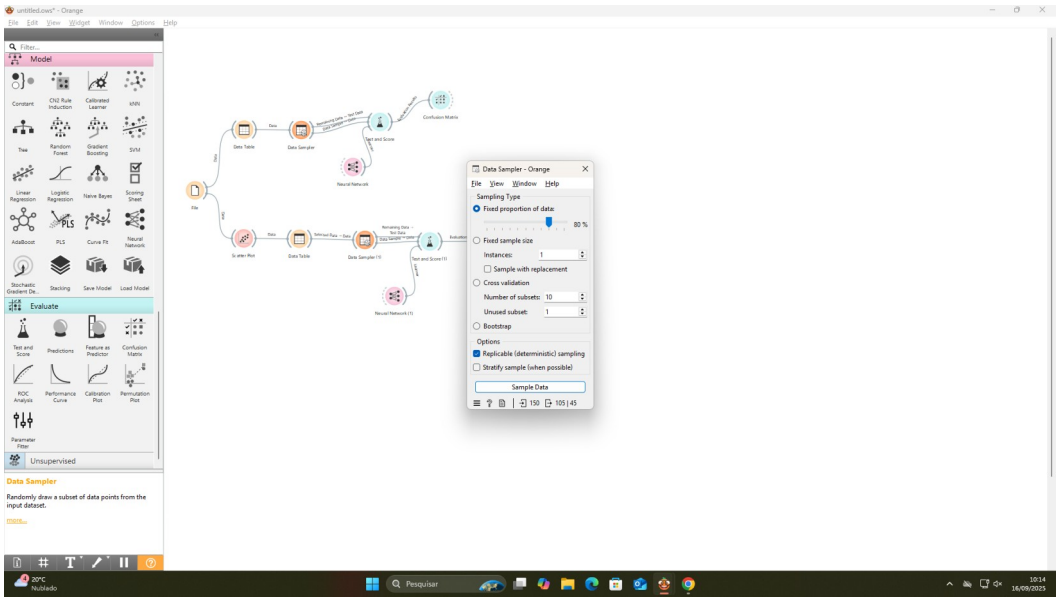
Figura 1



A Figura 1 ilustra o fluxo de trabalho completo do experimento no ambiente Orange, destacando a conexão entre os widgets de processamento de dados e o treinamento do modelo. A divisão inicial do dataset foi feita com 70% das amostras destinadas ao treinamento.

2. **Treinamento após pré-processamento**: As instâncias conflitantes ou outliers identificadas foram removidas do dataset. Em seguida, o modelo foi novamente treinado com a base de dados refinada, utilizando uma nova divisão de 80% para treinamento e 20% para teste.

Figura 2

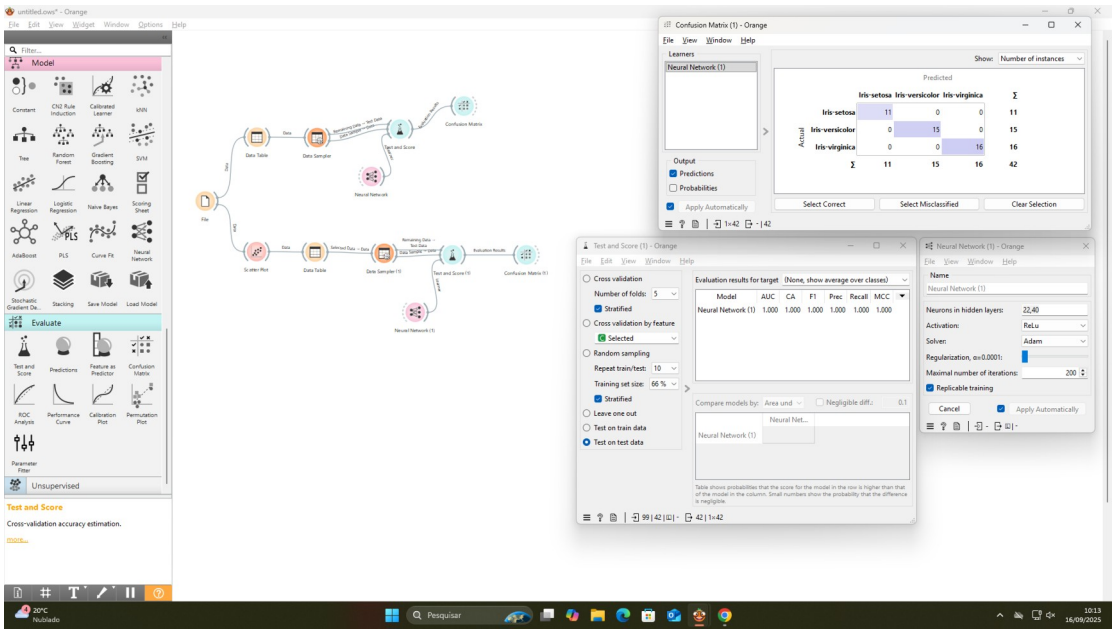


A Figura 2 mostra uma configuração alternativa para a divisão de dados, utilizando 85% para treinamento, o que pode ser explorado em futuras otimizações.

Análise dos resultados

No treinamento inicial, utilizando todos os dados, o modelo apresentou desempenho muito satisfatório, atingindo acurácia próxima de 100%. Após a exclusão das instâncias problemáticas identificadas, o modelo foi novamente treinado, mantendo resultados estáveis e até ligeiramente superiores. Isso demonstra que, embora o dataset Iris seja relativamente limpo, a remoção de casos problemáticos pode contribuir para uma maior robustez do modelo.

Figura 3



A Figura 3 apresenta a matriz de confusão e os resultados do modelo após o re-treinamento com a base de dados refinada. Os resultados confirmam que a exclusão de dados problemáticos contribuiu para a estabilidade e a robustez do modelo.

CONCLUSÃO

O trabalho demonstrou a importância do pré-processamento e da análise crítica dos dados em tarefas de classificação. A metodologia de duas fases, com a remoção de instâncias conflitantes ou anômalas e o posterior re-treinamento com uma divisão mais favorável (80% para treino), levou a resultados consistentes e robustos, confirmando a hipótese de que o tratamento adequado dos dados é fundamental para a qualidade dos modelos de aprendizado de máquina. Em aplicações futuras, recomenda-se ampliar a base de dados, aplicar técnicas de validação cruzada mais extensas e realizar ajustes de hiperparâmetros para maximizar o desempenho do modelo.

REFERÊNCIAS BIBLIOGRÁFICAS

ORANGE Data Mining Toolbox. Disponível em: <https://orange.biolab.si/>. Acesso em: set. 2025.

FISHER, R. A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, v. 7, p. 179-188, 1936.

Assinatura do orientador: _____

Assinatura do bolsista: _____