INF-0077 - T1 - Versionamento de Dados

Grupo:
Fabio Grassiotto
Wandemberg Santana Pharaoh Gibaut
Guilherme Ramirez

Link para o Github

O projeto se encontra hospedado no link https://github.com/fabiograssiotto/INF-0077

Decisões tomadas

Para este projeto utilizamos o DVC para o versionamento e gerenciamento dos dados. Para storage remoto o Google Drive foi utilizado devido à facilidade de utilização.

Comandos executados

Os comandos principais a seguir foram executados no terminal para configuração do DVC e versionamento dos dados e código do projeto:

```
git clone https://github.com/fabiograssiotto/INF-0077.git
pip install dvc-gdrive
dvc init
dvc remote add -d gdrive gdrive://1Pis_hAFuJADGfu5UJ7Ynt26jj51mDq7M -f
dvc add ds
dvc add Untitled-1.py
git add
git commit -m "Add DVC data"
git push origin main
dvc push
```

Screenshot do resultado do treinamento

```
[venv:mlops] (git:main) $ python Untitled-1.py

125/125 [==============================] - 229s 2s/step - loss: 0.7571 - accuracy: 0.5063 - val_loss: 0.6925
- val_accuracy: 0.5060
Epoch 2/10
125/125 [==============================] - 156s 1s/step - loss: 0.6901 - accuracy: 0.5336 - val_loss: 0.6792
- val_accuracy: 0.5645
Epoch 3/10
125/125 [==============================] - 211s 2s/step - loss: 0.6788 - accuracy: 0.5723 - val_loss: 0.6595
- val_accuracy: 0.6326
Epoch 4/10
125/125 [==============================] - 210s 2s/step - loss: 0.6657 - accuracy: 0.6020 - val_loss: 0.6588
- val_accuracy: 0.6084
Epoch 5/10
125/125 [==============================] - 154s 1s/step - loss: 0.6688 - accuracy: 0.5971 - val_loss: 0.6429
- val_accuracy: 0.6557
Epoch 6/10
125/125 [==============================] - 157s 1s/step - loss: 0.6544 - accuracy: 0.6211 - val_loss: 0.6421
- val_accuracy: 0.6411
Epoch 7/10
125/125 [==============================] - 215s 2s/step - loss: 0.6433 - accuracy: 0.6332 - val_loss: 0.6190
- val_accuracy: 0.6623
Epoch 8/10
125/125 [==============================] - 245s 2s/step - loss: 0.6328 - accuracy: 0.6491 - val_loss: 0.6053
- val_accuracy: 0.6764
Epoch 9/10
125/125 [==============================] - 306s 2s/step - loss: 0.6299 - accuracy: 0.6442 - val_loss: 0.6308
- val_accuracy: 0.6361
Epoch 10/10
125/125 [==============================] - 259s 2s/step - loss: 0.6240 - accuracy: 0.6495 - val_loss: 0.6175
- val_accuracy: 0.6497
[venv:mlops] (git:main) $
```

Conclusão

Pudemos notar na execução deste primeiro exercício que o **DVC** (Data Version Control) é uma ferramenta madura de gerenciamento de dados para machine learning. A instalação foi rápida em um ambiente padrão Python utilizando *virtual environments.* Os comandos utilizados para adição ao storage seguem o padrão do git, o que torna um primeiro aprendizado rápido e indolor.

A integração do DVC com o Google Drive é muito bem executada, sendo apenas necessária a aprovação da conta do Google em um ambiente de navegador de internet. Como todos os passos foram executados em uma máquina com sistema operacional Windows, não tivemos quaisquer problemas para executar o exercício.

Pudemos observar que o DVC se integra muito bem com o Git, substituindo os arquivos gerenciados em storage remoto por arquivos com links descritivos. Adicionalmente, a extensão disponível para o Visual Studio Code é muito eficiente para as atividades de adição e gerenciamento dos dados versionados.