Data Science do ZERO

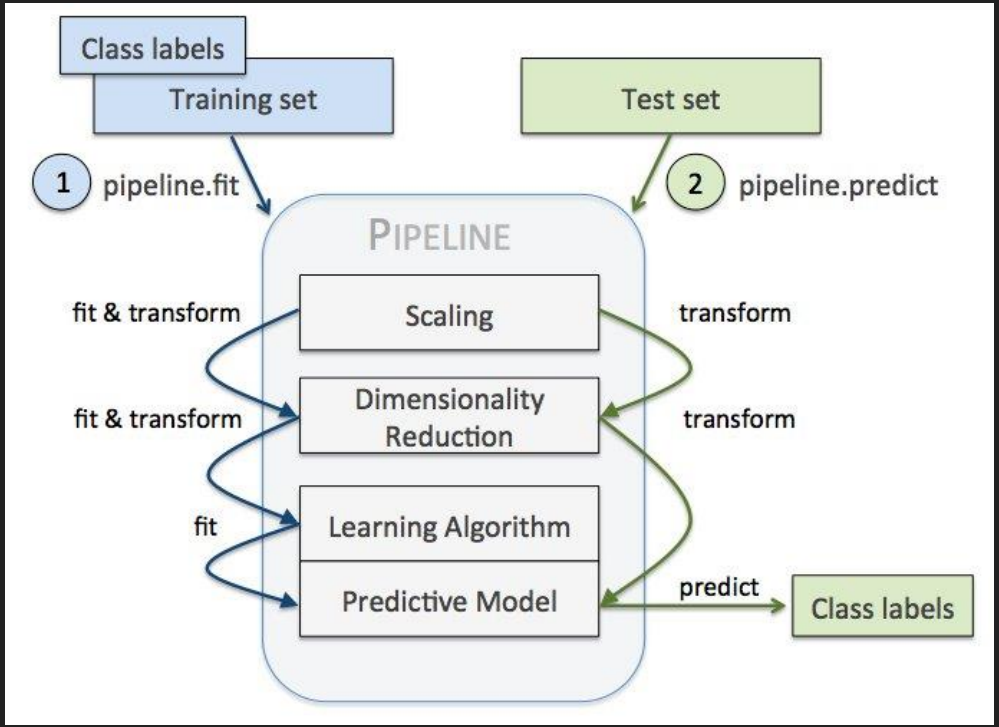
Pipelines

Pipelines

- Pipeline é um recurso útil para trabalhar com automação de processos em tarefas de Machine Learning.
- Esse recurso é usado para automatizar passos de forma sequencial.
- Permite a padronização e evita erros de manipulação de conjuntos de dados.

Pipelines

- Com pipelines garantimos que conjuntos de dados de treino e teste são processados de forma correta.
- Com a automatização dos fluxos evitamos erros de manipulação de dados e ajuda na reprodução de código.
- A aplicação sequencial das etapas dos dados de treino e teste da segurança e robustez nos processos.



Pipelines

- É possível criar diversos Pipelines com características distintas.
- Pipelines usando Scikit-learn permite encapsular etapa de pré-processamento de dados aplicação de um algoritmo de Machine Learning.
- É possível encapsular mais de um pré-processador.

```
pip_1 = Pipeline([
    ('scaler',StandardScaler()),
    ('clf', svm.SVC())
])

pip_2 = Pipeline([
    ('min_max_scaler', MinMaxScaler()),
    ('clf', svm.SVC())
])

pip_3 = Pipeline([
    ('scaler',StandardScaler()),
    ('pca', PCA(n_components=2)),
    ('clf', svm.SVC(kernel='rbf'))
])

pip_4 = Pipeline([
    ('scaler',StandardScaler()),
    ('clf', svm.SVC(kernel='poly'))
])

pip_5 = Pipeline([
    ('scaler',StandardScaler()),
    ('clf', svm.SVC(kernel='linear'))
])
```

Hands on!