DATALEARNS247

Machine Learning dengan PySpark

Introduction to MLlib

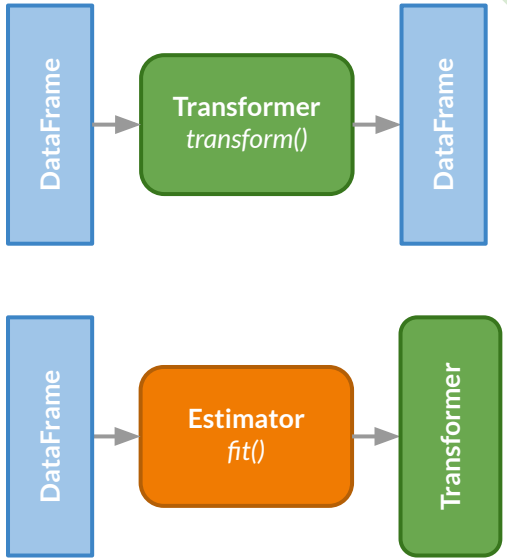
01 Sekilas MlLib

Sekilas MlLib

- MLlib adalah library machine learning Spark. Mencakup:
  - Algoritma: klasifikasi, regresi, clustering, dan kolaboratif filtering
  - Utilitas terkait fitur: ekstraksi fitur, transformasi, pengurangan dimensi, seleksi
  - Pipelines: tools untuk membuat, mengevaluasi, dan tuning workflow Machine Learning
  - Persistensi : menyimpan dan memuat kembali algoritma, model, serta pipeline
  - Utilitas: aljabar linier, statistik, pemrosesan data, dll.
- Mulai Spark 2.0, API utama untuk MlLib adalah DataFrame API
- Dokumentasi selengkapnya dapat dilihat di: https://spark.Apache.org/docs/latest/ml-guide.html

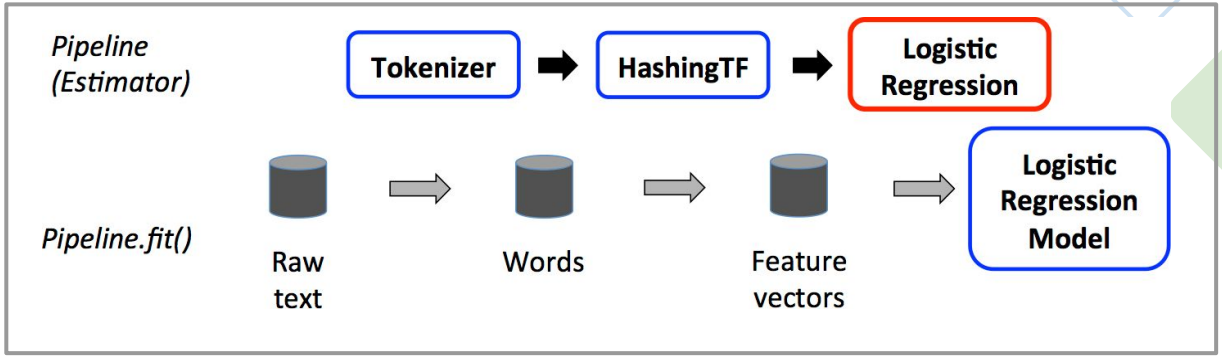
Pipelines

- API standar MLlib memudahkan untuk menggabungkan beberapa algoritma ke dalam satu pipeline
- Komponen utama dalam MlLib Pipelines adalah:
  - **Transformer**: mengubah satu DataFrame menjadi DataFrame lain.
  - **Estimator**: algoritma yang dapat dilatih (fitted) menggunakan DataFrame sebagai input. Hasilnya berupa Transformer.
- Pipeline menghubungkan beberapa Transformer dan/atau Estimator untuk membuat alur kerja/workflow ML

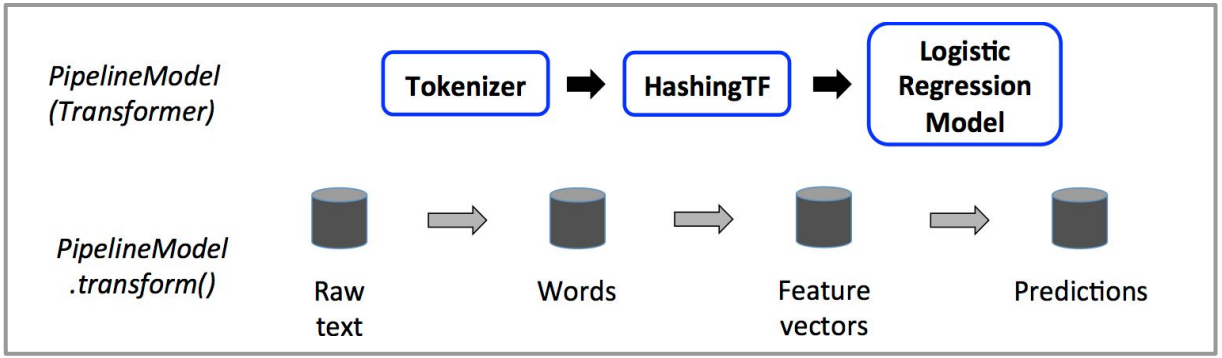


Contoh Pipelines

Training Pipeline



Testing Pipeline



https://spark.apache.org/docs/latest/ml-pipeline.html

Representasi Data

- Model pembelajaran mesin mengharuskan semua variabel input dan output berupa numerik
- Data kategorikal harus diubah menjadi numerik sebelum digunakan untuk training dan testing
- Beberapa metode pengkodean adalah: one-hot encoding, dummy encoding, hash, learned embedding, dll.
- Yang paling populer dan yang akan sering kita gunakan dalam pelatihan ini adalah : *one-hot* dan *dummy encoding*

name	qty
apple	5
banana	6
cherry	7
banana	8
apple	9
apple	5

One-Hot Encoded

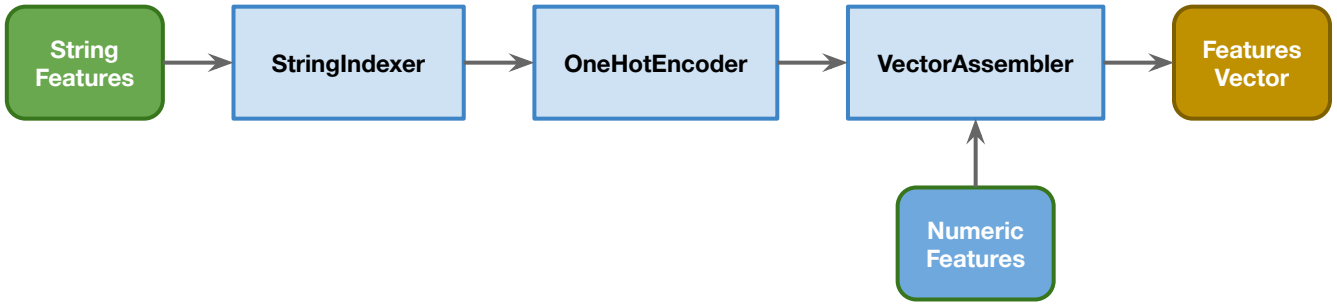
name	var1	var2	var3
apple	1	0	0
banana	0	1	0
cherry	0	0	1

var1	var2	var3	qty
1	0	0	5
0	1	0	6
0	0	1	7
0	1	0	8
1	0	0	9
1	0	0	5

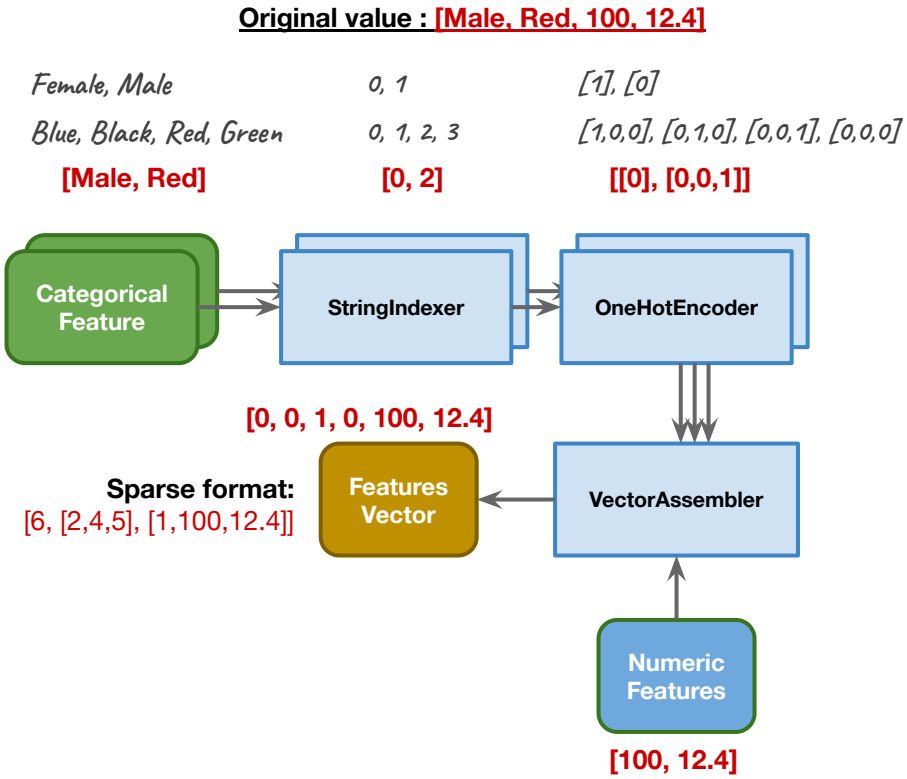
Pipeline Untuk Feature Preprocessing

Beberapa fitur transformer yang akan kita gunakan adalah:

- **StringIndexer** : estimator yang mengubah kolom string dari label menjadi indeks.
- **OneHotEncoder** : konversi fitur kategorikal menjadi vektor biner
- **VectorAssembler** : transformator yang menggabungkan sekumpulan kolom menjadi satu kolom vektor



Feature Preprocessing Pipeline in Action



Vektor feature direpresentasikan dalam *sparse format*, yaitu :

[Size, [Index of nonzero elements], [Values of nonzero element]]

Dense format	Sparse format
[0,7,3,0,0]	[5, [1,2],[7,3]]
[3,2,0,0,0,9,25,0,0,0,0,0,0,0,0]	[15, [0,1,6,7], [3,2,9,25]]
[0, 0, 1, 0, 100, 12.4]	[6, [2,4,5], [1,100,12.4]]

Labs 01

MlLib Introduction

MLlib Basics

Dalam Lab ini kita akan mempelajari tentang

- Contoh penggunaan MLlib, khususnya transformasi dan format data
- Bagaimana melakukan pemrosesan fitur menggunakan MLlib transformer dan membaca hasilnya:
  a. `StringIndexer`
  b. `OneHotEncoder`
  c. `VectorAssembler`
- Bagaimana menggunakan Pipeline

Labs 02

Regresi Linier dengan MlLib

MLlib Basics

- Dalam lab ini kita akan menjalankan regresi linier dengan data diabetes
- Kita akan menggunakan dataset yang sama dengan yang digunakan pada Regresi Linear dengan scikit-learn. Kali ini kita ambil data dari sumber asalnya, yaitu di sini : https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html
- Data yang kita gunakan adalah data yang sudah dinormalisasi, sama seperti data sample scikit-learn yang kita gunakan sebelumnya.

Labs 03

Decision Tree

Decision Tree & Random Forest

- Dalam labs ini kita akan melakukan klasifikasi dengan algoritma Decision Tree, menggunakan dataset subscriber churn