kod szkolenia: DBX-DE-ML / PL DL 2d

Databricks for Machine Learning – Data preparation

Dwudniowe szkolenie prowadzi uczestników przez cały proces przygotowania danych do uczenia maszynowego w środowisku Databricks – od wczytywania i przetwarzania danych (Delta Lake, DLT, Auto Loader) po zaawansowane przygotowanie cech, budowę pipeline’ów ML i rejestrację w MLflow i Feature Store.

Odbiorcy szkolenia

– Data Scientistci i ML Engineerowie
– Zespoły AI/ML i DataOps
– Inżynierowie danych przygotowujący dane do ML
– Uczestnicy z podstawową znajomością PySpark/ML

Korzyści

– Praktyczna znajomość pełnego procesu przygotowania danych do ML
– Budowa pipeline’ów z DLT, Auto Loader i Delta Lake
– Zastosowanie najlepszych praktyk: feature engineering, imputacja, encoding, standaryzacja
– Wersjonowanie i reużywalność danych z MLflow i Feature Store
– Praca w środowisku gotowym do wdrożeń produkcyjnych

Program szkolenia

 Dzień 1 – Data Engineering pod ML

1. Wprowadzenie do Databricks i workspace

2. Tworzenie i eksploracja DataFrame w SQL i PySpark
   • Tworzenie DataFrame, transformacje, proste agregacje
3. Wczytywanie danych z CSV/JSON/Parquet/Delta
4. Delta Lake: ACID, MERGE, UPDATE, DELETE, time travel
5. Batch and Streaming Load z Auto Loader
   • read/readStream, trigger once vs continuous, cloudFiles
6. Warstwowa struktura Medallion: Bronze → Silver → Gold
7. DLT: tworzenie pipeline'ów GUI, `CREATE LIVE TABLE`, triggery, expectations
8. Każdy moduł kończy się demem i ćwiczeniem praktycznym
 Dzień 2 – Przygotowanie danych do ML i pipeline'y modelowe
1. Databricks for ML & eksploracja cech (EDA)
   • MLflow, Feature Store, `display`, `summary`, `profile`
2. Fundamentals of Feature Engineering
   • Czym jest, extraction vs selection, agregacje, transformacje
3. Podział danych: train/test, cross-validation, stratified, sampling
4. Data Imputation
   • dropna, fillna, Imputer, flagi braków
5. Data Encoding
   • StringIndexer, OneHotEncoder, Target Encoding, wstęp do embeddingów
6. Data Standardization
   • StandardScaler, MinMaxScaler, RobustScaler – zastosowania i zalety
7. Budowa pipeline'u ML z MLflow
   • VectorAssembler, Pipeline, fit/transform, logowanie modeli, parametrów, metryk
8. Feature Store
   • Tworzenie, rejestracja, wersjonowanie i udostępnianie cech

Oczekiwane przygotowanie uczestnika

– Znajomość podstaw ML (regresja/klasyfikacja)
– Znajomość PySpark lub SQL
– Doświadczenie z pracą na danych lub pipeline'ach (zalecane)
Znajomość ML lub pipeline'ów przetwarzania danych (podstawowa).
Aby w pełni skorzystać z tego szkolenia, kluczowe jest dopasowanie poziomu kursu do obecnych umiejętności.

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii
Metoda szkolenia:
Wykład (25%),
demo i warsztaty (35%),
ćwiczenia praktyczne (40%)
Główne narzędzia: Databricks, Delta Lake, Auto Loader, DLT, MLflow, Feature Store

Język

Wykład: polski
Materiały: angielski

Metoda egzaminacyjna

Szkolenie nie zawiera formalnego egzaminu. Zaleca się jednak podejście do egzaminu Databricks Certified Machine Learning Professional jako kolejny krok rozwojowy.

Czas trwania

2 dni / 14 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Zaleca się jednak podejście do egzaminu Databricks Certified Machine Learning Professional jako kolejny krok rozwojowy.