



TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies ■

TECHNOLOGIE PRO BIG DATA

CVIČENÍ VII.

APACHE SPARK III

Lukáš Matějů
22.11.2023 | TPB





PŘÍPRAVA CVIČENÍ

- pro práci s MLLIB je vyžadován balíček numpy
 - bohužel není v kontejneru defaultně dostupný
 - zároveň je kontejner non-root, takže ho není možné doinstalovat

➤ potřeba upravit docker-compose.yml

- u mastera přidat řádek user: root
- potřeba znovu zavolat docker-compose up -d --build

```
services:
  spark:
    image: docker.io/bitnami/spark:3
    user: root
    environment:
```

➤ v kontejneru jsou nyní k dispozici root práva

➤ možnost nainstalovat numpy

- pip install numpy

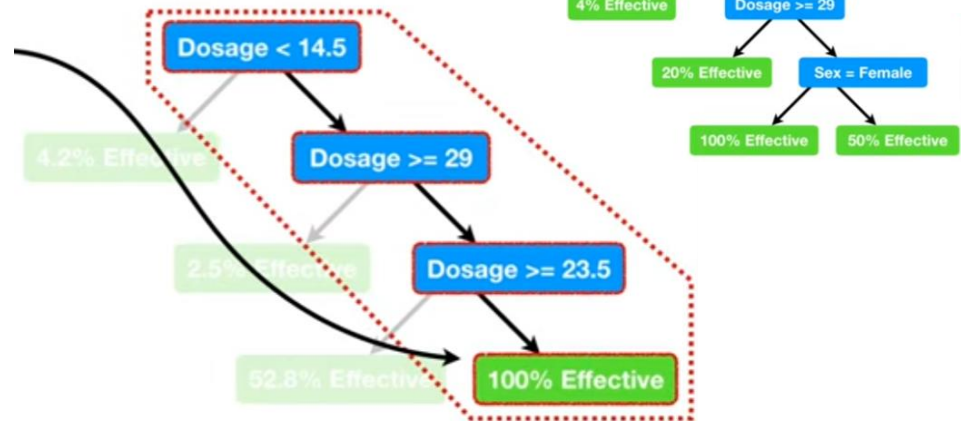
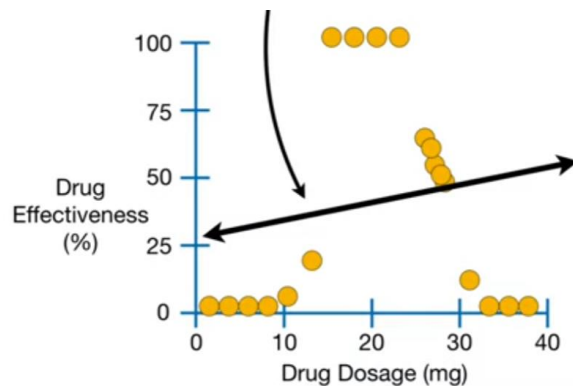
```
root@a42213733e4a:/files#
```

```
root@a42213733e4a:/files# pip install numpy
```

- poznámka: Spark bude opět ukecaný
 - pro ztišení je potřeba znovu vytvořit soubor log4j2.properties

ROZHODOVACÍ STROMY

- staví klasifikační nebo regresní modely ve formě stromů
 - dělí dataset na menší a menší části, zatímco budují přiřazený strom
 - finální strom má rozhodovací a koncové uzly
 - rozhodovací uzly mají dvě nebo více větví s jednotlivými možnostmi
 - koncové uzly reprezentují číselné hodnoty



<https://www.youtube.com/watch?v=g9c66TUylZ4>

DNEŠNÍ CVIČENÍ

1. predikujte hodnotu nemovitostí s využitím rozhodovacích stromů
 - k dispozici máte data o reálném prodeji nemovitostí na Tajvanu (realestate.csv)
 - datum transakce, stáří domu, vzdálenost k veřejné dopravě, počet obchodů, souřadnice a cena

TransactionDate	HouseAge	DistanceToMRT	NumberConvenienceStores	Latitude	Longitude	PriceOfUnitArea
2012.917	32	84.87882	10	24.98298	121.54024	37.9
2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
2013.583	13.3	561.9845	5	24.98746	121.54391	47.3

- jako příznaky použijte stáří domu, vzdálenost k veřejné dopravě a počet obchodů
- cílem je predikovat cenu (za jednotku)
- k implementaci
 - datový soubor má hlavičku pro snadné načtení do DataFrameu
 - pro použití více vstupních příznaků můžete použít VectorAssembler

```
assembler = VectorAssembler().setInputCols(["col1", "col2", ...])  
df = assembler.transform(data).select("labelColumnName",  
"features")
```

<https://www.udemy.com/course/taming-big-data-with-apache-spark-hands-on/>

DNEŠNÍ CVIČENÍ

1. predikujte hodnotu nemovitostí s využitím rozhodovacích stromů

- k implementaci
 - data rozdělte v poměru 90:10 na trénovací a testovací
 - pro odhad použijte rozhodovací stromy
 - k dispozici je DecisionTreeRegressor
 - není potřeba nastavovat hyper-parametry
 - setLabelCol slouží pro upřesnění sloupce se značkami
 - výsledky vypište ve formě predikovaná hodnota – reálná hodnota

```
(52.640625, 58.1)
(52.640625, 59.0)
(44.99000000000001, 59.6)
(67.7, 60.7)
(39.08484848484849, 61.5)
(52.640625, 62.1)
```

- jak kvalitní jsou výsledky?
- jak by je bylo možné vylepšit?
- **BONUS:** úlohu vyřešte na aktuálním cvičení