Spark – ML

MLlib jest biblioteką uczenia maszynowego dostępną w Sparku. Jej głównym celem jest udostępnienie praktycznych mechanizmów uczenia maszynowego w sposób skalowalny i prosty zarazem. Modele uczenia maszynowego wyliczane w oparciu o większe ilości danych mogą być bardziej dokładne, bardziej precyzyjnie oddając „kształty rzeczywistości". Dlatego też mechanizmy uczenia maszynowego, które w sposób skalowalny można przetwarzać w ramach klastrów na wolumenach skali Big Data zyskują na znaczeniu.

W ramach zadań, będziemy korzystali z biblioteki *MLlib*, która oparta jest na typach `DataFrame`. Znana jest ona także pod nazwą *ML,* w odróżnieniu od funkcjonującej w trybie wsparcia, biblioteki opartej na typach RDD (a mającej także nazwę *MLlib*).

Wykorzystamy dwa zbiory danych.

- Pierwszy z nich do implementacji klasycznego rozwiązania wykorzystującego metodę regresji liniowej… tak, cóż może być bardziej klasycznego jak wyznaczanie cen domów. `https://www.kaggle.com/mihirhalai/sydney-house-prices`
- Drugi ze zbiorów danych wykorzystamy do implementacji modelu klasyfikatora `https://www.kaggle.com/uciml/mushroom-classification`

Zawartość zbioru danych

Ceny domów w Sydney

Pierwszy ze zbiorów danych (`SydneyHousePrices.csv`) zawierający około 200 tysięcy cen domów w Sydney z lat 2000-2019 zawiera następujące kolumny:

- `date` – data sprzedaży
- `id` – zanonimizowany identyfikator nieruchomości
- `suburb` – dzielnica/przedmieście Sydney
- `postalCode` – kod pocztowy
- **`sellPrice`** – cena sprzedaży
- `bed` – liczba sypialni
- `bath` – liczba łazienek
- `car` – liczba miejsc dla samochodów
- `propType` – typ nieruchomości

Shrooming

Drugi ze zbiorów pochodzi oryginalnie ze repozytorium UCI Machine Learning (`https://archive.ics.uci.edu/ml/index.php`), który zawiera ponad pół tysiąca interesujących zbiorów danych przeznaczonych do implementacji przykładowych modeli uczenia maszynowego. Kolumny w tym zbiorze danych (`mushrooms.csv`) są następujące:

- **`class`** – określenie kategorii grzyba (jadalny (*edible*) = e, trujący (*poisonous*) = p)
- `cap-shape` – kształt kapelusza (bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s)
- `cap-surface` – powierzchnia kapelusza (fibrous=f, grooves=g, scaly=y, smooth=s)
- `cap-color` – kolor kapelusza (brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y)
- `bruises` - zasinienia (bruises=t, no=f)
- `odor` – zapach (almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s)

- `gill-attachment` – sposób doczepienia blaszek (spodu) grzyba do nóżki (attached=a, descending=d, free=f, notched=n)
- `gill-spacing` – przestrzeń pomiędzy blaszkami (close=c, crowded=w, distant=d)
- `gill-size` – rozmiar blaszek (broad=b,narrow=n)
- `gill-color` – kolor blaszek (black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y)
- `stalk-shape` – kształt nóżki (enlarging=e, tapering=t)
- `stalk-root` – korzeń (bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?)
- `stalk-surface-above-ring` – powierzchnia nóżki nad pierścieniem (fibrous=f, scaly=y, silky=k, smooth=s)
- `stalk-surface-below-ring` – powierzchnia nóżki pod pierścieniem (fibrous=f, scaly=y, silky=k, smooth=s)
- `stalk-color-above-ring` – kolor nóżki nad pierścieniem (brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y)
- `stalk-color-below-ring` – kolor nóżki pod pierścieniem (brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y)
- `veil-type` – typ osłony (partial=p, universal=u)
- `veil-color` – kolor osłony (brown=n, orange=o, white=w, yellow=y)
- `ring-number` – liczba pierścieni (none=n, one=o, two=t)
- `ring-type` – typ pierścienia (cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z)
- `spore-print-color` – kolor zarodników (black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y)
- `population` – populacja (abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y)
- `habitat` – siedlisko (grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d)

Aby zrozumieć wartości poszczególnych cech warto odwiedzić:
`https://biolwww.usask.ca/fungi/glossary.html`

Przygotowanie środowiska i zbioru danych

1. Uruchom klaster *Dataproc* – klaster Hadoop z dodatkowymi komponentami dostępny w ramach *Google Cloud Platform.*
   Korzystając z poniższego polecenia i konsoli *Cloud Shell* utwórz klaster.

```
gcloud dataproc clusters create ${CLUSTER_NAME} \
  --enable-component-gateway --region ${REGION} \
  --master-machine-type n1-standard-4 --master-boot-disk-size 50 \
  --num-workers 2 --worker-machine-type n1-standard-2 --worker-boot-disk-size 50 \
  --image-version 2.1-debian11 --optional-components JUPYTER \
  --project ${PROJECT_ID} --max-age=3h
```

2. Pobierz eksport notatnika `SP05_w1_23-Spark-ML-zadania.ipynb`

3. Na stronie środowiska *Jypyter* zaimportuj ten notatnik, a następnie otwórz zaimportowany notatnik. Znajdziesz tam dalsze instrukcje.