CHAPTER 2 • HANDS-ON MACHINE LEARNING

Տվյալների բեռնման ավտոմատացում

California Housing dataset-ի ներբեռնում, բացում և Pandas DataFrame-ի ստեղծում

Նպատակը և ելակետային վիճակը

Իրական Project-ներ

Տվյալները ստացվում են բազաներից, մի քանի աղյուսակներից կամ ֆայլերից:

Մեր Պարզեցված Դեպքը

Մեկ սեղմված ֆայլ՝ `housing.tgz`, որի ներսում `housing.csv`-ն է:

Անհրաժեշտ գրադարանների ներմուծում

Pathlib

Ֆայլերի և ուղիների հետ աշխատանք:

Pandas

CSV ֆայլի փոխակերպում DataFrame-ի:

Tarfile & Urllib

Արխիվների բացում և ինտերնետից բեռնում:

Քայլ 1–3. Ուղու սահմանում և ստուգում

Ֆայլի տեղադրություն

Ստեղծվում է `datasets/housing.tgz` ուղին պահպանման համար:

Կրկնակի բեռնման կանխում

Ստուգվում է՝ արդյոք ֆայլն առկա է. եթե այո, նորից չի ներբեռնվում:

Քայլ 4–5. Թղթապանակի ստեղծում և բեռնում

Թղթապանակի ապահովում

`mkdir(exist_ok=True)`-ն ստեղծում է `datasets/` շտեմարանը առանց սխալների:

GitHub-ից Ներբեռնում

`urllib.request.urlretrieve`-ը բեռնում է արխիվը ուղիղ հասցեից:

Քայլ 6. Արխիվի բացում (Extraction)

Tarfile գրադարան

Բացում է compressed `.tgz` ֆայլը և հանում պարունակությունը:

Անվտանգության ֆիլտր

`filter="data"` պարամետրն ապահովում է անվտանգ extract Python 3.12/3.13-ում:

Քայլ 7–8. Ընթերցում և DataFrame-ի ստացում

Pandas `read_csv`

CSV ֆայլից տվյալները վերածվում են հարմարավետ Pandas DataFrame-ի:

Արդյունքի կիրառություն

Հնարավորություն է տալիս անմիջապես օգտագործել `housing_full.head()` հրամանը:

Ամբողջական պրոցեսի հոսքը (Pipeline)

GitHub
Source URL

→

housing.tgz
Download

→

housing.csv
Extract

→

DataFrame
Pandas Read

ՀԻՄՆԱԿԱՆ ԳԱՂԱՓԱՐԸ

Ավտոմատացում՝ ամեն անգամ ձեռքով չկատարելու համար

Փոխանակ ամեն անգամ առանձին քաշելու, բացելու և կարդալու, մենք ստեղծում ենք մեկ ընդհանուր `load_housing_data()` ֆունկցիա, որն ամեն ինչ անում է ավտոմատ: