Report - 7th lesson

Data preprocessing — bu mashina o‘rganish yoki data analysis jarayonidagi muhim bosqichlardan biri bo‘lib, unda xom ma'lumotlar tozalab, analiz uchun tayyorlanadi. Ushbu darsda asosan missing values(yo‘qolgan qiymatlar) bilan ishlash va encodingusullari o‘rganildi.

Yo‘qolgan qiymatlar data sifatiga salbiy ta’sir ko‘rsatishi mumkin. Shuning uchun ularni to‘g‘ri boshqarish kerak. Darsda quyidagi usullar ko‘rib chiqildi:

isnull()yoki .isna()funksiyalari yordamida yo‘qolgan qiymatlar aniqlanadi.

Ularning soni va qaysi ustunlarda mavjudligi tahlil qilinadi.

Missing valuesni qanday to‘ldirish bo‘yicha 5xil usullarni o‘rgandik.

1. Drop missing values: bu yerda dropna() funksiyasi yordamida missing values mavjud bo‘lgan qatordagi ma’lumotlarni o‘chirish.

2. Fill with mean (ya’ni o‘rtacha qiymat bilan to‘ldirish): Sonli ustunlar uchun missing qiymatlarni ustunning o‘rtacha qiymati bilan to‘ldirish.

3. Fill with mean (o‘rtacha qiymati bilan to‘ldirish): Sonli ustunlar uchun missing qiymatlarni ustunning o‘rtacha qiymati bilan to‘ldirish.

4. Fill with mode (eng ko‘p uchraydigan qiymat bilan to‘ldirish):
Toifaviy (categorical) ustunlar uchun eng ko‘p uchraydigan qiymat (mode) bilan to‘ldirish.

5. Fill with a constant value:
Yo‘qolgan qiymatlarni maxsus belgilar yoki raqamlar bilan to‘ldirish (masalan, "Unknown", -1).

Undan tashqari for loop yordamida Missing values bilan ishlash.

Darsda for loop yordamida ma’lumotlar to‘plamidagi barcha ustunlar bo‘ylab

missing values aniqlanib, har bir ustun uchun kerakli to‘ldirish usuli avtomatik tarzda qo‘llanildi. Masalan:

```
for col in df.columns:
    if df[col].df.type == “object”:
        if df[col].nunique() <= 5:
            dummies = pd.get_dummies(df[col].prefix=col, dtype=int)
            df = pd.concat([df.drop(columns=col), dummies], axis=1)
else:
    df[col]=encoder.fill_transform(df[col])
```

Bu kod barcha ustunlardagi missing qiymatlarni to‘g‘ri usul bilan avtomatik to‘ldiradi.

3. Encoding (Kodlash) Usullari

Ma’lumotlardagi toifaviy (categorical) ma’lumotlarni modelga tushunarli qilish uchun sonli shaklga o‘tkazish kerak. Darsda quyidagi 5 ta encoding usullari ko‘rib chiqildi:

1. Label Encoding:
   Har bir toifa sonli qiymatga (0,1,2,...) aylantiriladi. Oddiy va tez, ammo kategoriyalar orasida sun’iy tartib bo‘lishi mumkin.
2. One-Hot Encoding:
   Har bir toifa uchun alohida binary ustun yaratiladi. Kategoriyalar orasida bog‘lanish bo‘lmaydi, lekin ustunlar soni ko‘payadi.
3. Ordinal Encoding:
   Kategoriyalar ma’lum tartibda sonlarga aylantiriladi, agar kategoriya tartibli bo‘lsa ishlatiladi.
4. Binary Encoding:

Kategoriyalarni binar kod shaklida ifodalaydi, bu ustunlar sonini kamaytiradi.

5. Frequency Encoding:
   Har bir kategoriya qanchalik ko'p uchrashi (freq) son bilan kodlanadi.

4. Xulosa

Dars davomida missing values bilan ishlashning bir nechta usullari va ularni avtomatik ravishda for loop yordamida qo'llash ko'rsatildi. Shuningdek, toifaviy ma'lumotlarni sonli formatga o'tkazish uchun asosiy 5 ta encoding usuli tushuntirildi. Ushbu usullar data sifatini yaxshilash va modellarni aniqroq ishlashi uchun juda muhim hisoblanadi.