**2-qism.8-dars**

Maqsad: Modelni o’qitish;

Train dataset Independent variables ( features → xususiyatlar) va depending variables (Target → maqsadli o’zgaruvchilarni) o’z ichiga oladi.

Ma’lumotlarni o’rganish orqali ma’lum

**Pattern (naqsh)**

- Pattern — bu **takrorlanadigan qonuniyat yoki bog‘liqlik** bo‘lib, ma'lumotlar ichida kuzatiladi. Bu ma'lumotlar ichidagi yashirin munosabatlarni ifodalaydi.
- **Odatda nimani anglatadi?**
    - Takroriy tendensiyalar yoki xarakterli xususiyatlar.
    - Naqsh ma'lumotlar asosida avtomatik ravishda ochiladi.
- **Misollar**:
    - Xaridlar ma'lumotida:
        - "Agar xaridor non sotib olsa, katta ehtimol bilan sut ham sotib oladi."
    - Tasvir tahlilida:
        - "Tasvirda dumaloq shakl va ikki chiziq bo‘lsa, bu ehtimol bilan ko‘z bo‘lishi mumkin."
- **Ma'lumotlar tahlilida roli**:
    - Bashorat qilishda ishlatiladi (masalan, tendensiyalar yoki takroriy holatlarni prognoz qilish).

**Framework (Tuzilma)**

- Framework — bu **tizimatik usul yoki qoidalar majmuasi** bo‘lib, muammolarni hal qilish yoki jarayonlarni tashkil etish uchun ishlatiladi.
- **Odatda nimani anglatadi?**
    - Modellar, algoritmlar yoki yondashuvlar to‘plami.
    - Tuzilma aniq qoidalar yoki jarayonlarni belgilaydi.
- **Misollar**:
    - Mashinani o‘qitishdagi frameworklar:

        - TensorFlow, PyTorch: Neyron tarmoqlarni yaratish va o‘qitish uchun platformalar.
    - Jarayon frameworklari:
        - CRISP-DM: Data mining jarayonlari uchun standart model.
- **Ma'lumotlar tahlilida roli**:
    - Naqshlarni aniqlash uchun ma'lumotlarni o‘rganish yoki modellashtirishni osonlashtiradi.

**EDE (Explotory Data Analysis)**

- Dastlabki ma’lumotlarni tahlil qilish jarayoni

**KLIB ???**

**KLIB** — bu Python uchun mo‘ljallangan kutubxona bo‘lib, u ma'lumotlarni tahlil qilish va ishlash jarayonlarini tezlashtirish va soddalashtirish uchun foydalaniladi. Ayniqsa, **EDA (Exploratory Data Analysis)** jarayonlarini avtomatlashtirish uchun qulay vosita hisoblanadi.

pip install klib

KLIB oddiy va tezkor EDA uchun bir nechta qulay funksiyalarni taqdim etadi.

- **Ma'lumotlar haqida umumiy ma'lumot olish:**

import klib import pandas as pd

**Ma'lumotlarni yuklash**

df = pd.read_csv("data.csv")

**Umumiy tozalash va statistik ma'lumotlar**

klib.clean_column_names(df) # Ustun nomlarini tozalash klib.describe(df) # Statistik xususiyatlar

klib.missingval_plot(df) # Yo‘q qiymatlar grafikasi

klib.corr_mat(df) # O‘zgaruvchilar orasidagi korrelatsiyani hisoblash klib.corr_plot(df) # Korrelatsiya matritsasining grafik ko‘rinishi

klib.cat_plot(df) # Kategorik ustunlarning taqsimoti klib.dist_plot(df['column_name']) # Sonli ustunning taqsimoti

| Xususiyat | KLIB | Seaborn | Matplotlib |
|---|---|---|---|
| Maqsadi | Avtomatlashtirilgan EDA | Statistik vizualizatsiyalar | Moslashtirilgan grafiklar |
| Foydalanish qulayligi | Juda oson | Oson | Murakkabroq |
| Grafik sozlash | Cheklangan | O‘rtacha | Juda moslashuvchan |
| Avtomatlashtirish | Ha | Yo‘q | Yo‘q |
| Moslik | Tezkor tahlil uchun | Statistik grafiklar uchun | Maxsus grafiklar uchun |
| Eng Ko‘p Foydalanadigan Kodek Misollari | klib.describe(df), klib.missingval_plot(df) | sns.heatmap(corr), sns.pairplot(df) | plt.plot(x, y), plt.scatter(x, y) |

**ROC and AUC**

ROC ( Receiver Operating Characteristic) Egri chizig’i.

ROC - Modelni qanchalik yaxshi ishlayotganini baholaymi. Bu grafik orqali modelingiz qanchalik yaxshi qarorlar chiqarayotganini baholaymiz.

**Misol:**

- **Tibbiyotda:** Kasallikni (masalan, diabetni) aniqlovchi model.
    - **True Positive Rate (TPR)**: Model kasallikni bor deb topdi va bu to‘g‘ri (masalan, kasallik haqiqatan bor edi, va model buni to‘g‘ri topdi).
    - **False Positive Rate (FPR)**: Model kasallik bor deb aytdi, lekin aslida yo‘q edi.

ROC egri chizig‘i — bu kasallik bor-yo‘qligini aniqlashdagi to‘g‘ri va xato qarorlar o‘rtasidagi muvozanatni ko‘rsatadi.

- **X o‘qi (FPR):** Xato topilgan holatlar.
- **Y o‘qi (TPR):** To‘g‘ri aniqlangan holatlar.

**AUC (Area Under the Curve)**

AUC — bu ROC egri chizig‘ining ostidagi maydonni o‘lchaydi va modelingiz umumiy ishlashini ko‘rsatadi.

**AUCni tushuntirish:**

- **Tibbiyotda:** Agar AUC = 1.0 bo‘lsa, model barcha kasallik holatlarini to‘g‘ri topmoqda (ideal model).
- **AUC 0.5 bo‘lsa:** Model tasodifiy taxmin qilayotgan bo‘lib, ishlashi yomon.
- **AUC > 0.8 bo‘lsa:** Model yaxshi ishlamoqda.

**Sohaga mos izoh:**

- **Sug‘urta:** Agar AUC yuqori bo‘lsa, modelingiz mijozning xavfini aniq baholay oladi.
- **Ta'lim:** Agar AUC yuqori bo‘lsa, test natijalariga qarab o‘quvchining bilim darajasini aniq topadi.

**Plotly Express - ma'lumotlarni interaktiv va chiroyli grafiklar yordamida tahlil qilish uchun qulay va oson foydalaniladigan vosita.**

pip install ipywidgets

pip install plotly bu ikkasi terminalga

import plotly.express as px bu pythonda

**Plotly Expressning asosiy vazifalari:**

1. **Interaktiv grafiklar yaratish:**
    - Grafiklarni avtomatik interaktiv qilish (masalan, sichqoncha bilan ustiga bosish yoki kattalashtirish imkoniyati).
2. **Ko‘p turdagi grafiklarni yaratish:**
    - **Chiziq grafiklar (line plot)**
    - **Scatter plot (nuqta grafik)**
    - **Histogram**
    - **Bar chart (ustun diagramma)**
    - **Pie chart (doira diagramma)** va boshqalar.
3. **Qulay va ixcham kod yozish:**
    - Murakkab grafiklar uchun oddiy bir qator kod kifoya.

4. **Interaktiv vizualizatsiya bilan ma'lumotlarni tahlil qilish:**
    - Ko‘p o‘zgaruvchi bilan ishlash va kategoriyalarni avtomatik ajratish.

**Smote ( Synthetic Minority Oversampling)**

**from imblearn.over_sampling import SMOTE** — bu **imbalanced-learn** (imblearn) kutubxonasidagi **SMOTE (Synthetic Minority Oversampling Technique)** algoritmini import qilish uchun ishlatiladigan kod. SMOTE **ma'lumotlar to'plamidagi balansni tiklash** uchun ishlatiladi.

**SMOTE nima?**

SMOTE — **kamchilikdagi sinflarni (minority class)** ortiqcha o'rnaklash (oversampling) orqali sinflar o'rtasidagi nomutanosiblikni hal qilish uchun ishlatiladigan usul.

**Masala:**

Ko'p real hayotdagi ma'lumotlar to'plamlarida sinflar balanssiz bo‘ladi. Masalan:

- **Tibbiyotda**: 95% sog‘lom va 5% kasallik holatlari.
- **Moliyaviy firibgarlik**: 99% odatiy tranzaksiyalar va 1% firibgarlik.

Nomutanosib ma'lumotlar to‘plami bilan ishlashda mashinani o‘qitish modellari **ko‘pchilik sinfga (majority class)** ko‘proq e'tibor berib, **kamchilik sinfni (minority class)** noto‘g‘ri klassifikatsiya qilishi mumkin.

**SMOTE qanday ishlaydi?**

1. Kamchilik sinfdagi mavjud o‘rnaklar o‘rtasida **sintetik (yangi) o‘rnaklar** yaratadi.
2. Bu o‘rnaklarni mavjud o‘rnaklar bilan aralashtirib, sinflarni balanslashga yordam beradi.

**KLIB**

**KLib** — bu Python kutubxonasi bo‘lib, u **ma'lumotlarni tozalash**, **EDA (Exploratory Data Analysis)**, va **vizualizatsiya** jarayonlarini soddalashtirish uchun mo‘ljallangan. Bu kutubxona yordamida ma'lumotlarni tahlil qilish va tayyorlash jarayonini tezlashtirishingiz mumkin.

| A | B | C | D |
|---|---|---|---|
| 0 | Ha | Qizil | 3.5 |
| 1 | Yo'q | Yashil | 2.3 |

| A | B | C | D |
|---|---|---|---|
| 1 | Ha | Qizil | 4.7 |
| 0 | Yo'q | Yashil | 1.1 |
| 1 | Ha | Ko'k | 0.8 |

Bu yerda:

- A: **Binary** (faqat 0 va 1 bor).
- B: **Matnli kategorik ma'lumot** (Ha/Yo'q).
- C: **Kategoriya (Ranglar)**.
- D: **Sonli ma'lumot**.

#Categorical data

binary = [col for col in df.columns if set(df[col].unique()) <= {0, 1}]

category = [col for col in df.select_dtypes(include = ['object','category']).columns]

category += binary

**Nega ustunlarni ajratamiz?**

- Sonli ustunlar (masalan, D) bilan **to'g'ridan-to'g'ri ishlay olamiz**, lekin:
    - **Matnli ustunlar (B, C) yoki Binary (A) ma'lumotlar**ni avval kodlashimiz kerak, chunki mashina faqat sonlarni tushunadi.

**Kod nima qiladi?**

1. **Binary ustunni topadi (A)**:
    - Tekshiradi: "Bu ustunda faqat 0 va 1 bormi?" Ha bo‘lsa, uni **binary** deb belgilaydi.
2. **Matnli kategorik ustunlarni topadi (B, C)**:
    - Tekshiradi: "Bu ustun matnlardan yoki kategoriyalardan iboratmi?" Ha bo‘lsa, uni kategorik ro‘yxatga qo‘shadi.
3. **Hammasini bir joyga yig‘adi**:
    - Binary (A) va matnli kategorik ustunlarni birlashtiradi, shunda ular bilan ishlash oson bo‘ladi.

---

**Natija qanday ko‘rinadi?**

- **Binary ustunlar:** ['A']
- **Kategorik ustunlar:** ['B', 'C', 'A']