MASHINALI O‘QITISHDA O‘RGATUVCHI TANLANMANI (DATASET) YARATISH USULLARI VA MATEMATIK MODELLARI

A.Mirzaqulov
Farg‘ona davlat universiteti fiz-mat. fan nomzodi, dotsent
R.Valixanov
Farg’ona davlat universiteti Amaliy matematika va informatika yo’nalishi II-bosqich magistri

Annotatsiya: *Ushbu maqolada mashinali o‘qitishda o‘rgatuvchi tanlanma (Dataset) va mashinali o‘qitishda tanlanmani yaratish usullariturlari haqida so‘z boradi. Buning uchun Ma’lumotlar to‘plami (Dataset) turlariga e’tibor qaratilib, Mashinali o‘qitishda tanlanmani yaratish usullari o‘rganib chiqildi.*

Kalit so‘zlar: *Ma’lumotlar to‘plami (Data set), ma’lumotlar bazasi, Data Table, Pandas,* ***Min,*** *Describe, dtypes, Python, Chiziqli diskriminant, logistik regressiya, model, Data Row va Data Column.*

Mashinali o‘qitish uchun ma’lumotlar to‘plami jadval shaklida tuzilgan va qayta ishlanadigan ma’lumotlardir [1]

Maqolada ma’lumotlar to‘plamni yaratish va yuklash, ma’lumotlar to‘plamini tahlil qilish usullarini, ma’lumotlarni vizualizatsiya qilish usullarini, mashinali o‘qitish algoritmlarini baholash, ma’lumotlarni bashorat qilish usullari o‘rganilgan.

Ushbu maqolada mavjud .csv fayli asosida DataSetni qanday yaratishni, ma’lumotlar to‘plamini qayta ichlashni, matematik asoslarini va mashinali o‘qitishda foydalanish usullari o‘rganilgan.

Ma’lumotlar to‘plamidagi qiymatlarga asoslanib, ba’zi mezonlar bo‘yicha talabalarning o‘zlashtirish natijalarini aniqlash, dasturiy ta’minotini o‘rgatish va ta’limni amalga oshirish vazifasi bayon qilinadi.

Ixtiyoriy ma’lumotlar to‘plamini yuklash uchun pandas modulidan foydalanamiz. Shuningdek, tavsiflovchi statistika va ma’lumotlarni vizualizatsiya qilish uchun ma’lumotlarni o‘rganish uchun pandasdan foydalanamiz.

Pandas modulidagi read_csv funksiyasi .csv faylini o‘qish uchun tayinlangan.

```
pandas as pd
df=pd.read_csv('d:\\namuna.csv')
df
```

	familiya	ism	informatka	differ	tarmoqlar	baza_dannix	oper_tizim	umumiy	ortacha
0	soliev	farhod	4	4	4	3	5	4	4
1	karimov	islom	3	4	5	4	5	3	4
2	azimov	ali	5	4	3	4	4	5	4
3	olimov	ahad	3	3	4	4	5	4	4
4	oripov	salim	4	4	5	5	3	4	4
5	valiev	ikrom	4	4	5	5	3	3	4

Pandas kutubxonasidagi Min() funktsiyasi DataSet ning har bir ustuni uchun minimal qiymat chegarasiga ega bo‘lish imkonini beradi.

```
familiya      azimov
ism             ahad
informatka         3
differ             3
tarmoqlar          3
baza_dannix        3
oper_tizim         3
umumiy             3
ortacha            4
```

Berilgan jadvalda talabalarning fanlardan olgan baholari to‘g‘risidagi ma’lumotlar mavlud.

Namuna.csv faildagi ma’lumotlar to‘plami yuqoridagi dastur fragmenti yordamida df o‘zgaruvchiga o‘zlashtirib olingan.

Dastlab, Pandasda berilgan ma’lumotlar to‘plamin tahlil qilinadi.

Ma’lumotlar to‘plamidagi qatorlar va ustunlar sonini .shape usuli yordamida topish mumkin:

df.shape

(6, 9)

df.shape[0]

6

df.shape[1]

9

Ma’lumotlar to‘plamining xususiyatlari va statistik xarakteristikalari haqida asosiy tasavvurga ega bo‘lish uchun ushbu buyruq kifoya qiladi.

df.describe() buyruq berilsa ekranda ma’lumotlar to‘plamining muhim xarakteristik parametrlarini aniqlash mumkin bo‘ladi.

	informatka	differ	tarmoqlar	baza_dannix	oper_tizim	umumiy	ortacha
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.0
mean	3.833333	3.833333	4.333333	4.166667	4.166667	3.833333	4.0
std	0.752773	0.408248	0.816497	0.752773	0.983192	0.752773	0.0
min	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.0
25%	3.250000	4.000000	4.000000	4.000000	3.250000	3.250000	4.0
50%	4.000000	4.000000	4.500000	4.000000	4.500000	4.000000	4.0
75%	4.000000	4.000000	5.000000	4.750000	5.000000	4.000000	4.0
max	5.000000	4.000000	5.000000	5.000000	5.000000	5.000000	4.0

Ma’lumotlar to‘plamidagi zaruriy ustunlardagi ma’lumotlarni ekranga chiqarish amaliy ahamiyatga ega bo‘lib uni quydagi usul bilan tashkiletish mumkin.

df[['informatka’,'differ']]

	informatka	differ
0	4	4
1	3	4
2	5	4
3	3	3
4	4	4
5	4	4

Ma’lumotlar to‘plamidagi ustunlarda turli arifmetik amallarni bajarish namunasi quyida keltirilgan.

df['ortacha']=(df['informatka']+df['differ']+df['tarmoqlar']+df['baza_dannix']+df['oper_tizim'])/5

display(df)

	familiya	ism	informatka	differ	tarmoqlar	baza_dannix	oper_tizim	umumiy	ortacha
0	soliev	farhod	4	4	4	3	5	4	4.0
1	karimov	islom	3	4	5	4	5	3	4.2
2	azimov	ali	5	4	3	4	4	5	4.0
3	olimov	ahad	3	3	4	4	5	4	3.8
4	oripov	salim	4	4	5	5	3	4	4.2
5	valiev	ikrom	4	4	5	5	3	3	4.2

Ma’lumotlar to‘plamidagi indexli ustunni yaratish uchun quyidagi buyruqdan foydalaniladi. df.index=["a","b","c","d"]

Ma’lumotlar to‘plamidagi ustunlar ko‘p bo‘lsa, ularning ayrimlarini ekranga chiqarish uchun print(df.loc["c":]) buyruq o‘rinli. Bu yerda “c” malumotlar to‘plamidagi “c” indexli qator ekranga chiqarilgan bo‘ladi.

Ma’lumotlar to‘plamidagi har bir ustunlardagi ma’lumotlar turlicha bo‘lishi mumkin. Buning uchun quyidagi buyruqdan foydalaniladi.

df.dtypes

```
familiya        object
ism             object
informatka       int64
differ           int64
tarmoqlar        int64
baza_dannix      int64
oper_tizim       int64
umumiy           int64
ortacha        float64
```

df.dtypes dan ko‘rinadiki, ma’lumotlar to‘plami ikkita asosiy komponentdan iborat: ob’yekt, o‘zgaruvchining tipi o‘rganishda ma’lumotlar to‘plamidagi ob’yektlar qanchalik ko‘p bo‘lsa, u haqiqatga shunchalik yaqin bo‘ladi.

Ma’lumotlar to‘plamidagi maydonlarning tiplari qanchalik ko‘p bo‘lsa, tahlil qilish shunchalik qiyin bo‘ladi. Funktsiyalar sonining ko‘payishi bilan ma’lumotlar to‘plamini qayta ishlashning murakkabligi chiziqli emas, balki eksponent, ya’ni juda tez o‘sadi.

Keltirilgan namuna an’anaviy ma’lumotlar bolib, bir kompyuterdan boshqariladigan ma’lumotlar bazalarida tuziladi va saqlanadi, ya’ni, raqamli yoki matn qiymatlarini o‘z ichiga olgan jadval ko‘rinishidir.

Ma’lumotlar to‘plamining jadvalidai berilganlar tasodifiy sonlardan iborat bo‘lsa, ma’lumotlar to‘plamimi qayta ishlashda ehtimollik model tanlanadi. Ehtimoliy ma’lumotlarni yaratish modeli ma’lumotlar to‘plamidagi parametrlar tasodifiy shakllantirilishini nazarda tutadi.

Ma’lumotlar to‘plamining jadvalidagi berilganlarni o‘zaro bog‘likligiga asoslanib logistik regressiya modeli, chiziqli diskriminant tahlili modeli, parametrlarlarni o‘lchash aniqligiga ko‘ra keng yaqin qo‘shnilar modeli tanlanadi.

Agar uning barcha elementlari mustaqil ravishda taqsimlangan bo‘lsa, tanlama oddiy deb ataladi. Oddiy namuna bir qator mustaqil tajribalarning matematik modeli bo‘lib, odatda mashinani o‘rganish uchun ishlatiladi. Shu bilan birga, Machine Learningning har bir bosqichi o‘ziga xos ma’lumotlar to‘plamini talab qiladi [3].

Ma’lumotlar to‘plamini pythonda turli usullar bilan yaratib olish mumkin, jumladan, DataFrame jadvalli ma’lumotlar strukturasidan foydalanish mumkin. Har bir jadvalda har doim qatorlar va ustunlar mavjud.

Misol sifatida Python lug‘ati yordamida DataFrame yaratish eng oson:

import pandas as pd

df = pd.DataFrame({'uzbekcha': ['kitob', 'qalam', 'uy', 'gul'],'Rusch': ['kniga', 'karandash','dom','roza'],

'inglizcha': ['book','pencil', 'house', 'flower']})

df

	Rusch	inglizcha	uzbekcha
0	kniga	book	kitob
1	karandash	pencil	qalam
2	dom	house	uy
3	roza	flower	gul

:

DataFrame yordamida hosil qilingan jadvalni .csv fayl sifatida saqlash mumkin va ma'lumotlar to'plamiga o'zgartiriladi .csv faylni o'qib qayta ishlash imkoni mavjud bo'ladi. Ular quyidagi buyruqlar bilan amalga oshiriladi.

df.to_csv('d:\\fname.csv')

df = pd.read_csv('d:\\fname.csv', sep=',')

df

Natijada fname.csv fayl tarkibidagi ma'lumotlar to'plami df o'zgaruvchiga o'zlashtirib olingan bo'ladi. Ma'lumotlar to'plamidagi ixtiyoriy ustundagi ma'lumotlarni o'suvchi yoki kamayuvch tartibda sarlash uchun quyidagi buyruqdan foydalanish o'rinli

df.sort_values('uzbekcha', ascending=True).head()

Ma'lumotlar to'plamidagi ixtiyoriy ustundagi ma'lumotga nisbatan filtirlash uchun quyidagi namunadan foydalanish etarli

df[df['uzbekcha'] == 'gul']

Pythonda ma'lumotlar to'plami (dataset) ning matematik modelini shakillantirish uchun:

- Pandas kutubxonasi yordamida o'quv ma'lumotlar to'plami yaratailadi.
- Olingan ma'lumotlar to'plamiga mos matematik model o'rnatiladi.
- Bashorat qilish uchun Python kodini yoziladi.

Mashinali o'qitishda ma'lumotlar to'plamiga asoslanib o'quv ma'lumotlarini tayyorlash, fanlardan olgan baholari bo'yicha vedomostlar tayyorlash, semestrlarga mos holda kurs va guruhlarni shakillantirish zarur bo'ladi. Takidlangan muammolarni hal qilishda ma'lumotlar to'plami dataset oddiy yozuvlardan emas, balki tranzaksiyali, matrisali va grafli ma'lumotlardan iborat bo'ladi.

Xususiy holda matematika – informatika fakultetida 900 ta talabaning ma'lumotlar to'plami yaratilgan bo'lsa guruhni shakillantirish uchun ulardan 20-30 tasi tanlab olinadi.

Agar talabaning barcha parametrlar to'plami bir xil mos holdagi ustunlarda ifodalangan bo'lsa, masalaning yechimi sodda bo'ladi.

Hozirgi paytda turli sohalar uchun ma'lumotlar to'plami yaratilgan va qayta ishlanmoqda.

FOYDALANILGAN ADABIYOTLAR:

1. http://datareview.info/article/universalnyj-podxod-pochti-k-lyuboj-zadache-mashinnogo-obucheniya/
2. Introduction to Machine Learning with Python, by Sarah Guido, Andreas Müller (notebooks available here).
3. http://www.machinelearning.ru/wiki/index.php?title=Выборка
4. В.В.Вьюгин «МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ МАШИННОГО ОБУЧЕНИЯ И ПРОГНОЗИРОВАНИЯ» МОСКВА 2013